Formation and Signatures of the First Stars and Quasars

A thesis presented

by

Zoltán Haiman

to

The Department of Astronomy

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Astronomy

Harvard University
Cambridge, Massachusetts

May 1998

Abstract

Although direct detection of the very first stars and quasars in the universe is beyond the capabilities of current instruments, there is substantial evidence for the existence of these objects. Observations of the intergalactic medium (IGM) reveal that it was highly ionized, and enriched with heavy elements, before redshift $z \sim 5$. Standard cold dark matter (CDM) cosmological models predict that the first clouds indeed condensed prior to this redshift, around $10 \lesssim z \lesssim 30$. In this thesis, we study several questions related to the formation and observational signatures of the first generation of astrophysical objects in CDM cosmologies. Using one–dimensional simulations, we study the effects of gas pressure and H_2 cooling on the sizes of the first collapsed systems. We find that the H_2 abundance reaches high enough values to permit efficient cooling and collapse in the central, shock–heated regions of clouds with virial temperatures $T_{\rm vir} \gtrsim 100°$K. Still smaller clouds ($\sim 10^3 M_\odot$) are able to collapse by $z \sim 10$, due to the gravitational potential of collisionless dark matter. We then examine the effect of ultraviolet (UV) photons on the H_2 abundance in these clouds, and find that H_2 is universally photodissociated by the low UV flux from the first few stars. This imposes a high minimum mass ($\sim 10^8 M_\odot$) for typical halos that can cool and harbor astrophysical objects. Using phenomenologically calibrated semi–analytical methods, we calculate several indirect consequences of the first stars, including the reionization history of the IGM and the resulting suppression of the cosmic microwave background (CMB) anisotropies; or the optical depth of dust from high redshift Type I supernovae and the resulting spectral distortion of the CMB. Both of these effects will be observable with forthcoming satellites, such as MAP or Planck, while the Next Generation Space Telescope could directly image numerous $z > 10$ star–clusters. Similarly, we extrapolate the observed quasar population to $z \gtrsim 5$, and demonstrate that early quasars could have comparable effects. Finally, we show that data from the Hubble Deep Field already places useful constraints on these models.

To the memory of my grandfather
Haiman György

Acknowledgments

I would like to express my gratitude to all the people who have contributed to this thesis, scientifically or otherwise. I thank my advisor, Avi Loeb, for defining clear and interesting problems, and providing both excellent technical advice and a great deal of attention and enthusiasm. I thank Martin Rees for advising me during my year in England, and for his continuous insightful suggestions afterwards.

I thank Alex Dalgarno for his invaluable help with molecular chemistry; George Field for his generous advice and guidance; George Rybicki and Giovanni Fazio for serving on my TAC, and for their useful comments and suggestions for improvement. I thank my other collaborators: Anne Thoul for providing the 1–D hydro code, and Piero Madau for providing his expertise on the Hubble Deep Field.

I thank Margaret Geller and Ed Turner for being on my defense committe, and for their stimulating comments; and Jean Collins and Christina Doyle for making the administrative aspects of this thesis seem very smooth, and for their overall devotion to the graduate students.

I am grateful to my friends Joe Mohr and Danny Chung for sharing with me their fascination with science; Rohan Mahadevan for his incredible energy and enthusiasm; David Heyrovsky and Eliot Quataert for stimulating scientific and everyday conversations; my long time friends Gergő, Gergő and Zsolt for their friendship and unfailing interest. Finally, I could not have done this work without the love and support from my wife Latha, my parents, my brother Tamás, and my family in Hungary, the USA, and India.

I gratefully acknowledge partial financial support for this work from a Cambridge European Trust ORS award, the Isaac Newton Studentship in Cambridge, U.K., and a Harvard Merit Fellowship.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The universe as we observe it today has a remarkable hierarchical structure: the stars are arranged into galaxies, the galaxies into clusters of galaxies, and the clusters of galaxies into superclusters. Yet we infer from observations of the cosmic microwave background radiation (CMB), that the early universe was nearly perfectly smooth, and devoid of any significant structure. One would naturally like to know how the smooth and dense early universe evolved into its present–day form, containing a rich and regular structure. A particularly interesting epoch is the transition when the very first generation of astrophysical objects formed, lighting up the universe for the first time since the big bang, and polluting it with their nucleosynthesis products.

Although these object are beyond the limits of direct detection by present–day astronomical instruments, they are ideally suited for a theoretical exploration. Indeed, the predictive power of the standard cosmological models are the highest when the first astrophysical objects form, since the initial conditions in the early universe are well–defined, and the first objects do not yet affect each other through interactions and various feedback processes. In this thesis, we study when and how the first generation of astrophysical objects formed, and some of their expected properties. In addition to these theoretical predictions, we address the question of whether we could observe these objects in the foreseeable future, or at least infer their existence indirectly from their effects on their surroundings.

1.1 Theory of Formation of the First Objects

It has long been realized that the formation of structure in the universe is driven by gravitational instability (Gamow 1948, Novikov 1964; Peebles 1967). As early as 1933, Lemaitre applied Birkhoff's (1923) theorem to point out a simple property of a spherically symmetric perturbation in a homogeneous universe. A spherical region with average density ρ follows the same evolution as a homogeneous universe with the density ρ. Stated more generally, this picture implies that an initially slightly overdense region always decelerates relative to the overall expansion of the universe, eventually stops expanding, turns around and finally recollapses. It is natural to associate the end–product of such a collapse with some type of an astrophysical object, whose properties are determined by parameters such as the size, density, shape, angular momentum, and environment of the initial seed region. This association is hindered by the fact that in an expanding universe, initial inhomogeneities are amplified only as a power law in time, rather than the exponential amplification achieved by instabilities encountered in other areas of physics (Lifshitz 1946, Bonnor 1957). It was initially believed that the relatively low efficiency of gravitational instability rules it out as the process responsible for structure formation, until it was proposed that the universe could have had sufficiently large primordial inhomogeneities (Zel'dovich 1965; Peebles 1968).

In principle, given the full initial spatial density $\rho(\vec{r})$ and velocity $\vec{v}(\vec{r})$ fields at some early epoch, including any small fluctuations, the equations of gravitational instability can be used to compute the subsequent time–evolution of these quantities. Although this can be done exactly only in the linear regime (Peebles 1980), a general important feature of gravitational instability is that, in the absence of other forces, it has no intrinsic scale. This implies that two primordial seed regions with equal initial overdensities evolve together and collapse simultaneously, irrespective of their sizes (Fillmore & Goldreich 1984; Bertschinger 1985). However, surveys of galaxies reveal that the universe is nearly homogeneous on scales above a few hundred Mpc, while the departure from homogeneity is a strong function of scale on smaller scales (Geller & Huchra 1989). Although gas pressure introduces a characteristic minimum scale to gravitational instability (Jeans 1928), during the relevant phases of the structure formation epoch, this scale is $\sim 10^{5-6}\mathrm{M}_{\odot}$ (Peebles 1965) 5–6 orders of magnitude below galactic scales. The scale–dependence of the observed hierarchy of structure must instead be explained by the nature of the initial conditions.

It is often assumed that the initial density fluctuations obey random Gaussian statistics. While there is no definite fundamental theory that derives the initial density field, inflationary theories (Guth 1981; Linde 1982) predict this property. The assumption of Gaussianity can also be broadly justified by the central limit theorem (Bardeen et al. 1986). The principal advantage of Gaussian density fluctuation is it can be fully described by their power spectrum, $|\delta_k|^2$. The primordial power spectrum is usually assumed to have no preferred scale, $|\delta_k|^2 \propto k^n$. For constant–curvature fluctuations, $n = 1$ (Harrison 1970; Zel'dovich 1972), a result reproduced by standard inflationary theories (Steinhardt & Turner 1984; although more recent versions of inflation allow $n \neq 1$, Liddle & Lyth 1992). During the early, radiation–dominated stages of the universe, non–gravitational effects modify the shape of the power spectrum (Peebles 1993). The initial conditions around galactic scales are therefore most conveniently specified just after the recombination era, when gravity is indeed the dominant force on these scales, and structure formation begins. A particularly intuitive quantity is the root–mean–square mass fluctuation on a given mass scale M, given by $\sigma_M \propto \int dk k^2 |\delta_k|^2 \propto M^{-(n+3)/6}$.

The observed large–scale distribution of galaxies (Geller & Huchra 1989), the abundance of galaxy clusters clusters (Viana & Liddle 1996), and the temperature fluctuation of the CMB (Smoot et al. 1992; Bennett et al. 1996; Fixsen et al. 1996) can be used to place constraints on σ_M at the recombination epoch on a range of mass–scales. Although these constraints are not yet very precise, a generic conclusion supported by the existing data is that the initial density fluctuations were decreasing as a function of scale (Kolb & Turner 1990). This is indeed the case in the general class of models in which cold dark matter (CDM) dominates the dynamics of the universe, and processes the initial $n = 1$ power spectrum to predict $n = 1$ on large scales and $n = -3$ on small scales (Peebles 1982). In these models, the fluctuations at the beginning of the structure formation era decrease with M ($\sigma_M \propto M^{-2/3}$ for $M \gg 10^{12}\mathrm{M}_{\odot}$, while the dependence is only logarithmic for $M \ll 10^{12}\mathrm{M}_{\odot}$). Since the more overdense a region, the earlier it collapses, these types of fluctuations imply that

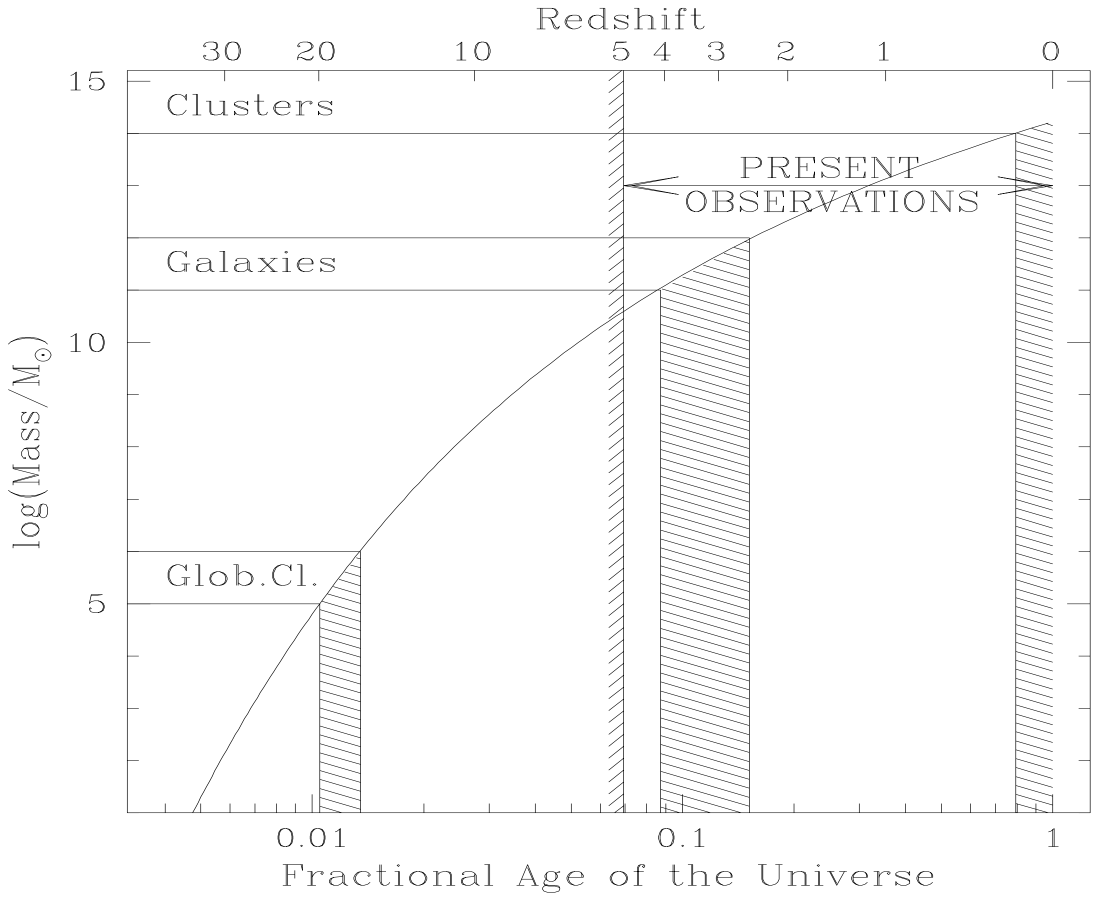


Figure 1.1: *The formation epoch of objects with various masses in a standard cosmological model (a CDM model with* $\Omega = n = \sigma_{8h^{-1}} = 1$ *and* $\mathrm{H}_0 = 75$ km s^{-1} Mpc^{-1}*), shown as a fraction of the present age of the universe. The upper labels indicate redshifts. Although present observations of the brightest quasars and galaxies probe the last 90% of the cosmic time, current instruments are not yet sensitive enough to image the first objects. The Next Generation Space Telescope[2] the successor to the Hubble Space Telescope, could make this detection.*

the present structure was built from the bottom up, with smaller objects appearing first, and subsequently merging or clustering together to assemble the larger objects (Peebles 1980). The predicted formation epochs of objects with various masses in the so–called standard CDM cosmology (Bardeen et al. 1986) are shown in Figure 1.1. Galaxies, which have masses around $10^{11-12}\mathrm{M}_\odot$, are expected to have formed when the universe had approximately 10% of its present age (redshift $z \sim 3$), just around the limit of the deepest present–day observations (i.e. the Hubble Deep Field, HDF; Williams et al. 1996). Clusters of galaxies with masses around $10^{14-15}\mathrm{M}_\odot$ are predicted to have formed as recently as 80% of the current age, with the more massive clusters still assembling at the present time. Objects with the masses of globular clusters, $10^{5-6}\mathrm{M}_\odot$, are predicted to have formed as early as $\sim$1% of the current age, or $z \sim 20$.

Considering only gravitational forces, the CDM model in Figure 1.1 would predict that still smaller objects formed at even earlier times. However, inside clouds with masses below the cosmological Jeans mass after recombination, $10^{5-6}M_\odot$ (Peebles 1965), the pressure gradient forces are comparable to gravity, and significantly alter the collapse dynamics for the baryonic gas. This leads to a complication in predicting the sizes and formation times of the first objects, since one must take into account gas pressure. Pressure depends on the various astrophysical processes that effect the gas temperature, including the radiative mechanisms both inside the cloud, such as H and He line cooling, and cooling by Bremmstrahlung, as well as the interaction of the cloud with the CMB via Compton scattering. In addition, Saslaw & Zipoy (1967) and Peebles and Dicke (1968) first pointed out that molecular hydrogen (H_2) is likely to be the most efficient radiative cooling mechanism inside the first objects; the chemical composition of the clouds must therefore be followed by including a realistic reaction network for H_2.

Putting these issues aside, another question concerns the detailed behaviour of each cloud during the late stages of collapse. Formally, in the absence of non–gravitational forces, a perfectly spherical top–hat perturbation simply collapses to a point (Peebles 1980). According to more accurate treatments describing self–similar solutions of spherical, but inhomogeneous secondary infall for a mixture of cold dark matter and baryons (Gunn & Gott 1972, Fillmore & Goldreich 1984, Bertschinger 1985), the evolution is as follows. In the initial stages, while the bulk of the cloud is still turning around or expanding, the central, densest region of the cloud collapses and forms a dense virialized clump. The dark matter component virializes through violent relaxation (Lynden–Bell 1967), while the kinetic energy of the gas is converted into thermal energy through a hydrodynamic shock that raises the gas temperature to its virial value. As subsequent gas shells fall towards the center, they encounter an outward–propagating shock, and are brought to a sudden a halt. As this process develops, material accretes onto the center, gradually establishing a stationary, virialized object with a $\rho \propto r^{-2.25}$ density profile. Early discussions of the formation of galaxies and clusters have argued that the subsequent behaviour of the gas in such a virialized object is determined by its ability to cool radiatively on a dynamical time (Rees & Ostriker 1977; Dekel & Silk 1986). The same ideas apply on the smaller scales expected for the first collapsed clouds (Silk 1977; Kashlinsky & Rees 1983). Objects that are unable to cool and radiate away their thermal energy maintain their pressure support and identity, until they become part of a larger object via accretion or mergers. On the other hand, objects that can radiate efficiently will cool and continue collapsing.

What happens next to a cloud that can cool efficiently? The most important parameters controlling the fate of these clouds are probably the angular momentum and the star formation rate (Padmanabhan 1993). If star formation proceeds efficiently on a cooling time, the gas turns into stars before it can contract by dissipating its energy. In this case, a roughly spherical stellar system would form, resembling an elliptical galaxy. On the other hand, if star–formation is inefficient, the cold baryons collapse on a free–fall time to form a thin disk, perpendicular to the angular momentum vector (Larson 1975). In this case, the evolution of the gas in the disk is still uncertain, but is likely controlled by its angular momentum. In hierarchical scenarios, the angular momenta of virialized clouds originate from tidal torques of neighbouring density perturbations, and have typical values of $\lambda \sim 5\%$ of the fraction of rotational support (Barnes & Efstathiou 1987). This angular momentum is still $\sim 10^3$ times higher than what would allow the gas to sink to the central Schwarzschild radius, and the gas is therefore expected to fragment and form an extended stellar system resembling a disk galaxy (Loeb & Rasio 1994). However, this can not always be the case. The integrated light of observed quasars implies that a non–negligible fraction ($\lesssim 3\times10^{-5}$) of all the baryons in the universe have been converted into $\sim 10^{6-10}M_\odot$

black holes (Soltan 1982), suggesting that the angular momentum barrier can somehow be overcome (Loeb 1993; Eisenstein & Loeb 1995b). The implication is that, in at least a fraction of the disk systems, the gas is able to collapse to form a massive central black hole. Conceivably, a small fraction of the gas ends up as a black hole in all systems (Haehnelt & Rees 1993). Subsequent accretion of gas onto the black hole would result in a luminous object resembling a quasar (Zel'dovich & Novikov 1964; Salpeter 1964).

1.2 Observational Evidence for Early Objects

From the observational point of view, the formation epoch of the first objects can be approached from two directions. On one hand, we can directly observe structures such as quasars, galaxies, and absorption systems as far back as $\sim 10\%$ of the present age of the universe (or $z \sim 5$, see Figure 1.1). The record holder quasar is currently at $z = 4.89$ (Schneider, Schmidt & Gunn 1991), and the highest redshift known for a galaxy is $z = 4.92$ (Franx et al. 1997), or possibly as high as $z = 5.34$ (Dey et al. 1998). Recent deep observations with the Hubble Space Telescope (HST; Steidel et al. 1996; Madau et al. 1996) and ground–based telescopes, such as Keck (Lowenthal et al. 1996), have provided an abundance of data on early galaxies, and considerably constrained their evolution and stellar content at $z \lesssim 5$. Similarly, observations of quasar spectra with high resolution spectrographs, both on HST and Keck, have revealed a multitude of absorption systems along the lines of sights to quasars at $z \lesssim 5$, allowing a detailed analysis of their properties.

One the other hand, observations of the CMB yield information on the small initial density fluctuations at the recombination epoch at $z \sim 1000$, only a few hundred thousand years after the big bang. The original detection of the CMB temperature anisotropies by the COBE satellite (Smoot et al. 1992) confirmed the notion that the present structure in the universe originated from small density fluctuations in the universe. Subsequent data allowed accurate constraints to be placed on the amplitude of the primordial density fluctuations (Bennett et al. 1996; Fixsen et al. 1996).

Despite these significant recent advances, the epoch when the first astrophysical objects started to appear, $z \sim 10 - 20$, around 1% of the present cosmic time, has not yet been probed directly. The Next Generation Space Telescope (NGST, see Smith & Koratkar 1998), scheduled for launch in the year 2007, will be able to probe this unobserved interval, and detect objects as small as $\sim 10^6 \mathrm{M}_\odot$ out to redshifts as high as $z \sim 10$, with its proposed ~ 1 nJy sensitivity. However, the ability of the first light–sources to affect their surroundings suggests that even if they are too faint to be imaged, their existence can be inferred indirectly. As first pointed out by Doroshkevich (1967), and later emphasized by Couchman & Rees (1986), nuclear burning in stars, or thin–disk accretion onto a Schwarzschild black hole in quasars both release $\gtrsim 10^6$ eV per hydrogen atom; however, the ionization of a hydrogen atom requires only ~ 10 eV. It is therefore sufficient to convert a fraction of $\sim 10^{-5}$ of the gas into either stars or black holes, in order to ionize the rest of the universe, suggesting that even if small and rare, the first objects could have had a significant effect on the ionization and temperature of the intergalactic medium (IGM). In addition, the first objects are expected to pollute the IGM with their heavy elements (Truran & Cameron 1971), and could also explain the missing mass in clusters of galaxies and galactic halos (White & Rees 1978). Subject to the uncertainties in the initial mass function (IMF), these effects are closely related to each other (Carr, Bond & Arnett

1984). A number of recent observational clues along these lines already constitute strong evidence for an early population of stars that formed at $z \gtrsim 5$.

1.2.1 Ionization of the Intergalactic Medium

The spectra of high–redshift quasars can be used as a sensitive probe of the amount of neutral hydrogen (column density) along the line of sight to the quasar. For any reasonable background gas mass density, even if a fraction as little as $\sim 10^{-5}$ of the gas between us and the quasar were in the form of neutral hydrogen, it would cause an easily measurable suppression (by an optical depth of $\tau \sim 10^6$) of the quasar flux at wavelengths shorter than the local Lyα resonance at the quasar's redshift (the so–called "Gunn–Peterson trough"; Gunn & Peterson 1965). Of the several hundred quasars with known spectra at $z \lesssim 5$, none has revealed such a suppression, implying that the gas in the IGM is highly ionized out to this redshift. Several authors have noted that radiation from the observed, declining bright quasar population at $z \gtrsim 3$ is insufficient to reionize the universe, and additional sources of ionizing photons are necessary (Shapiro & Giroux 1987; Miralda-Escudé & Ostriker 1990, Vikhlinin 1995).

1.2.2 Metal Enrichment of Lyman α Absorbers

Instead of the expected Gunn–Peterson trough, the spectra of high–redshift quasars revealed a multitude of absorption features, caused by discrete clouds between us and the quasar. These absorbing clouds, collectively called the Lyα forest because of the ubiquity of the Lyman α absorption lines, have neutral hydrogen column densities between 10^{13} and 10^{22} cm^{-2}. Recent high–resolution spectroscopic observations of the Lyα forest at $z \sim 3$ show absorption lines of various heavy elements, such as C and Si, indicating a $\sim 1\%$ solar abundance for these elements, in absorbers with HI column densities as low as 10^{14} cm^{-2} (Cowie 1996; Songaila & Cowie 1996; Tytler et al. 1995). Numerical simulations identify such absorbers with mildly overdense regions in the IGM (Hellsten et al. 1997, and references therein), whose low density makes them unlikely sites for any star–forming activity. Higher column density "damped" Lyα absorbers, which are thought to be the progenitors of present–day galaxies, show heavy element abundances at a similar level during the early phase of their formation, at $3.5 \lesssim z \lesssim 4.5$ (Lu et al. 1996). The universality of the mean metal abundance in these absorbers, spanning a range of several orders of magnitude in HI column density, and the presence of heavy elements even in the most tenuous systems at the highest redshifts, imply that an early phase of roughly homogeneous pre–enrichment by stars occurred throughout the universe at $z \gtrsim 5$.

1.2.3 Smoothing of Microwave Background Anisotropies

Observations of the anisotropy of the CMB temperature can be used to set additional constraints on the ionization history of the IGM. In particular, ionized gas interacts with the CMB via Thompson scattering, and reduces the anisotropies on angular scales below $\lesssim 10°$, the size of the horizon at the reionization redshift (Efstathiou & Bond 1987; Kamionkowski, Spergel, & Sugiyama 1994; Hu & White 1997). Recent ground–based microwave anisotropy experiments detected tentative evidence for excess anisotropies at these angular scales (Scott, Silk & White 1995; Bond 1995b), showing that the IGM was still neutral at redshifts $z \gtrsim 100$. Future satellite experiments, such as MAP or Planck, will be able to detect reionization at redshifts as low as $z \sim 10$ (Zaldarriaga, Spergel, & Seljak 1997). Putting all existing data together, the implication is that some, yet unidentified sources reionized the IGM during the redshift interval $5 \lesssim z \lesssim 100$.

1.2.4 Massive Compact Halo Objects

Finally, the formation of the first stars may be related to the well–known dark matter problem (White & Rees 1978). An old population of pre–galactic stars would behave as particles of a collisionless gas during the subsequent formation of galaxies, and populate the diffuse halos of these galaxies. In an effort to search for the constituent particles of the the dark halo of the Milky Way Galaxy, recent microlensing surveys have found that, subject to uncertainties of the detailed models of the spatial density distributions of the components of our galaxy, $\sim 50\%$ of the halo mass might be in the form of Massive Compact Halo Objects (MACHOs) with typical masses around $\sim 0.5 M_{\odot}$ (Alcock et al. 1997a,b; Flynn, Gould & Bahcall 1996). It remains to be seen whether future microlensing studies confirm these early reports; however, a stellar population that forms before galaxies are assembled could provide a natural explanation for these microlensing events.

1.3 Modeling the Formation of Early Objects

The conceptually most straightforward approach to study structure formation is to employ 3–dimensional numerical simulations. Given the initial conditions, and solving the relevant equations of gas dynamics, radiative transfer, and chemistry, a 3–dimensional calculation would, in principle, yield the full non–linear evolution of the density field. In practice, a complete treatment is beyond the capabilities of current computers and N–body and hydrodynamics codes. Three dimensional simulations have been restricted to study questions that involve a limited range of mass scales (Weinberg, Hernquist & Katz 1997), such as the formation of the Lyα forest (Hernquist et al. 1996), galaxy mergers (Dubinski, Mihos & Hernquist 1996), or the formation and structure of dark matter halos (Gelb & Bertschinger 1994; Navarro, Frenk & White 1997, 1998). While simulations are adequate to explore these questions, they are less well suited to study the first objects, where gas pressure can not be ignored, and a wide range of scales ($\sim 10^5 - 10^{14} \mathrm{M}_{\odot}$) must be resolved. Although recently there have been attempts to apply numerical simulations to the first structures, these calculations are just beginning to resolve the Jeans mass (Anninos et al. 1997; Gnedin & Ostriker 1997).

An alternative approach is to use semi–analytical methods, which have the general advantage of providing computationally inexpensive and accurate solutions. A frequently employed method that predicts the mass function of collapsed objects as a function of redshift was developed by Press & Schechter (1974), utilizing analytical solutions of spherically symmetric top hat collapse (see Peebles 1980), and the Gaussian random nature of the initial density fluctuations. Although this method does not treat gas pressure, unlike numerical simulations, it avoids the restriction on the range of mass scales by drawing a direct correspondence between the nonlinear mass function, and the initial density field. Since its appearance, numerous papers have clarified and amended the Press–Schechter theory (Peacock & Heavens 1990; Bond

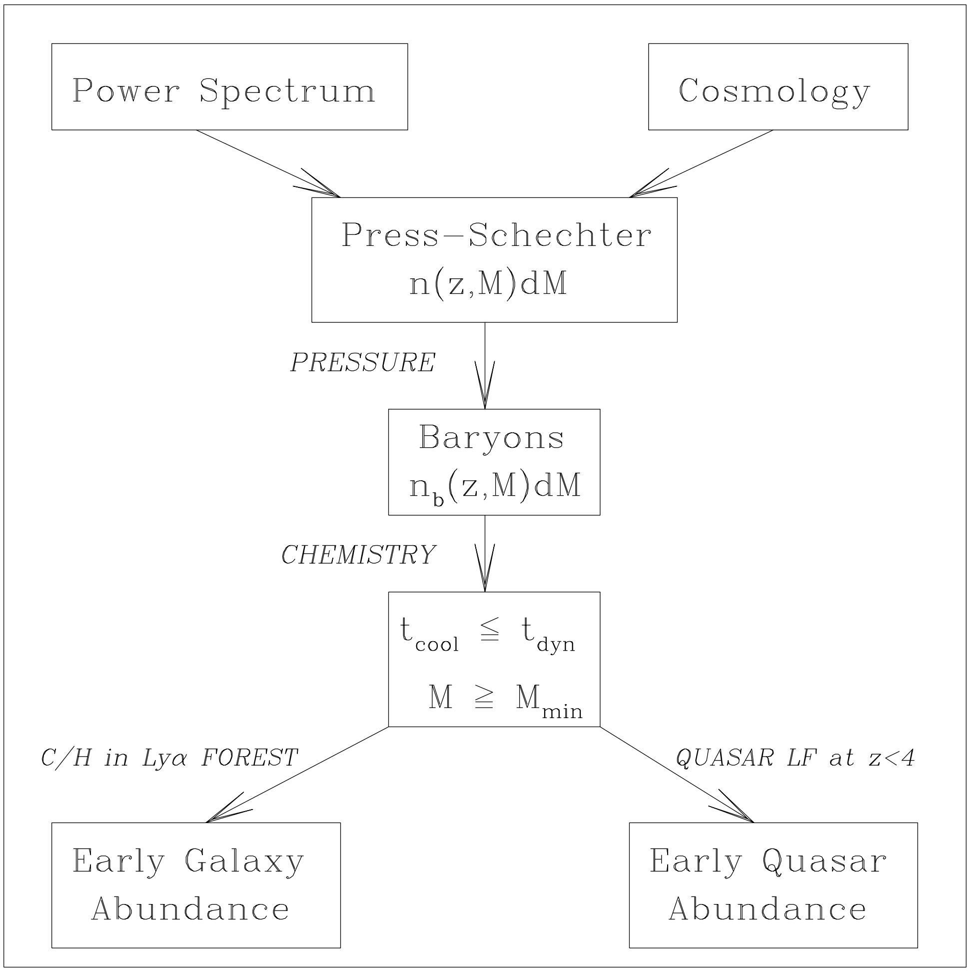


Figure 1.2: *A schematic picture of the ingredients of our model.*

et al. 1991; Lacey & Cole 1993, 1994; Kauffmann & White 1993), a comprehensive account of these efforts is given in Padmanabhan (1993). Although this theory does not correctly predict the masses and locations of individual objects, the predicted mass function, which only relies on the theory statistically, has been tested and found to agree well with numerical simulations (Katz et al. 1993; Klypin et al. 1993).

The Press–Schechter theory forms the back–bone of our modeling, as depicted schematically in Figure 1.2. Given the power spectrum at recombination and the cosmological parameters, the theory gives a simple expression for the mass function of dark matter halos at each redshift. Although collisionless dark matter collapses to form bound halos on all scales, the fraction of gas retained in each collapsed halo decreases below unity on scales below the Jeans mass, $\sim 10^6 M_\odot$. This results in a suppression of the baryonic mass function below this scale. In close analogy with the pressureless Press–Schechter theory, the suppressed mass function can be computed, given the evolution of a spherical gas cloud whose pressure is not negligible. Although no analytical solution is available for this case, the evolution can be obtained using spherically symmetric, one–dimensional simulations (Thoul & Weinberg 1995), which are still far less computationally expensive than their 3–dimensional counterparts.

As described in § 1.1, a necessary condition for each collapsed gas cloud to form any astrophysical object is that the gas can cool efficiently. Because of the chemical simplicity of the primordial gas, this requirement turns out to be rather restrictive (Rees & Ostriker 1977; Silk 1977). The main parameter that controls the cooling efficiency is temperature. Above $T \gtrsim 10^6$K, the most important mechanism is Bremsstrahlung; however, such high temperatures are above the virial temperatures in the relative shallow potential wells of the first objects. Between $10^4\mathrm{K} \lesssim T \lesssim 10^6\mathrm{K}$, cooling is possible via collisional excitation of neutral H and He. At lower temperatures, cooling relies on the presence of molecules with lower excitational energies; in the primordial gas the only molecule that could have a sufficient abundance is H_2, allowing cooling between $10^2\mathrm{K} \lesssim T \lesssim 10^4\mathrm{K}$ (see Dalgarno & Lepp for a review of astrochemistry in the early universe). Below $T \lesssim 10^2$K, the collapsed clouds are unable to cool, and remain pressure–supported for longer than a Hubble time. Because of its unique ability to cool the smallest (and therefore earliest) clouds, the abundance of H_2 can control the minimum sizes and formation times of the very first systems. In our models, we therefore calculate the abundance of H_2 via the full network of chemical reactions.

As described in § 1.1, systems that can cool efficiently lose their thermal pressure, and a fraction of the gas forms a central black hole, and/or fragments into a population of stars. Although various authors have attempted to model these processes (Hirasawa 1969; Matsuda, Sato & Takeda 1969; Larson 1974; Hutchins 1976; Silk 1977; Carlberg 1981; Palla, Salpeter & Stahler 1983; Lin & Murray 1992), a reliable estimate for the efficiencies are difficult to obtain from a fundamental theory. Given the large uncertainties, we have chosen a phenomenological approach, in which these efficiencies are adjusted by hand to match observational data. Similar approaches have been used in the past by Fukugita & Kawasaki (1993), Madau & Shull (1994), and Tegmark & Silk (1994). The efficiency of early star formation can be calibrated based on the observed heavy element content of the intergalactic medium. As outlined in § 1.2, the interpretation of the C and Si absorption detected in Lyα forest clouds $z \sim 3$, is that these heavy elements were produced by an early generation of stars prior to this redshift. The observations further suggest that pre–enrichment is roughly uniform across the IGM. Since the heavy element yields of stars are reliably known (Renzini & Voli 1981; Maeder 1992; Woosley & Weaver 1995), this interpretation uniquely normalizes the star formation efficiency, subject only to uncertainties in the shape of the IMF (see Larson 1986a,b for a discussion of the IMF at early times).

The black hole formation efficiency can be calibrated using data on the observed quasar population at $z \lesssim 4$. The number of quasars known (Hartwick & Schade 1990, Warren, Hewett & Osmer 1994) are sufficiently large so that the quasar luminosity function (LF), together with its evolution, can be constructed out to this redshift. Fitting formulae for the evolution of the quasar LF between $0 < z < 4.5$ have been obtained by Boyle et al (1991), and Pei (1995). The quasar LF is much shallower than the luminosity function of galaxies, and shows that the abundance of the brightest quasars peaks around $z \sim 2.5$, and declines rapidly at lower redshifts. There is additional evidence of a less rapid decline towards higher redshifts. Assuming a relation between halo mass and black hole mass, and a model for the black hole luminosity (Frenk, King & Raine 1992), the average quasar black hole efficiency in Press–Schechter type models can be adjusted to fit the quasar luminosity function in the observed redshift range (Small & Blandford 1992; Haehnelt & Rees 1993). Within a handful of galactic nuclei in the local universe, the black hole formation efficiencies can also be measured directly. In these galaxies, the black hole masses are typically a fraction $\sim 2 \times 10^{-3}$ of the bulge mass of their host (Kormendy et al. 1997; Magorrian et al. 1998). This should be representative of the black hole to halo mass ratio of high redshift quasars at the end of their active phase.

1.4 Thesis Outline

The work presented in Chapters 2–8 of this thesis deals with several questions related to the formation and observational signatures of the first objects that were motivated both by the theoretical considerations and the observational clues outlined above. Much of this work was done within the general framework summarized in § 1.3, and depicted schematically in Figure 1.2.

In Chapter 2, we compute the collapse of spherically symmetric clouds, treating collisionless dark matter, baryonic gas and H_2 chemistry with a 1–D code. We derive the minimum masses of clouds in which the H_2 abundance reaches sufficiently high values to effect the dynamics, and conclude that H_2 cooling is important only in the central, shock–heated regions. We demonstrate that primordial clouds as small as $10^3 M_\odot$ are nevertheless able to collapse at high redshifts, due to the gravitational potential of the cold dark matter component.

In Chapters 3 and 4, we examine the effect of ultraviolet photons on the H_2 abundance in homogeneous clouds. We conclude that a small amount of ultraviolet cosmic background flux, produced by the very first handful of stars, can photodissociate the H_2 molecules even in large clouds, thus eliminating the most important cooling agent for clouds with masses below $\sim 10^8 M_\odot$. On the other hand, we show that inside very dense objects, UV photons can have the opposite effect, and trigger an increase in the H_2 abundance, leading to a positive feedback between cooling and star–formation.

In Chapters 5 and 6, we compute several possible observable signatures of the early generation of stars. These include the suppression of the CMB anisotropies the stars produce by reionizing the IGM, the probability of stellar microlensing towards high redshift sources, the cumulative Bremsstrahlung emission from early star clusters, as well as the expected observational number counts of these clusters. We also compute the optical depth for absorption by dust produced in high redshift Type I supernovae, and the spectral distortion of the CMB due to absorption and re–emission of starlight by this dust.

In Chapter 7 and 8, we address the formation and signatures of high redshift quasars. We extrapolate the population of quasars observed at $z \lesssim 5$ by a physically justified scheme that fits the evolution of the observed quasar luminosity function around $z \sim 3$. We repeat the calculations of Chapters 5 and 6 for the quasar population, and conclude that the effects of quasars, in general, are comparable to those expected of stars. We confront the quasar formation model with the deepest available data, the Hubble Deep Field, and show that this data can be used to place useful constraints on the models.

Chapter 2

Cosmological Formation of Low-Mass Objects

Zoltán Haiman[1,3], Anne A. Thoul[2], and Abraham Loeb[1,2]

[published in *The Astrophysical Journal*, **464**, 523 (1996)]

ABSTRACT

We investigate the early formation of bound objects with masses comparable to the cosmological Jeans mass ($\sim 10^5 M_\odot$). We follow the growth of isolated spherically symmetric density peaks starting from the linear perturbative regime. The initial parameters correspond to density peaks of various widths and heights in a cold dark matter cosmology. We use a one-dimensional spherical Lagrangian hydrodynamics code to follow the dynamical, thermal, and non-equilibrium chemical evolution of the gas. The system includes a collisionless dark matter component and a baryonic component composed of the nine species $\mathrm{H}, \mathrm{H}^-, \mathrm{H}^+, \mathrm{He}, \mathrm{He}^+, \mathrm{He}^{++}, \mathrm{H}_2, \mathrm{H}_2^+$, and e^-. All relevant chemical reactions between these species and their cooling mechanisms are included in the calculations.

We explore the dependence of the dynamical evolution of the gas on two parameters: the initial mass scale and the initial overdensity of the system. We follow the evolution of the density, temperature, and abundance profiles within the cloud, assuming two types of central boundary conditions for the collisionless component: in one the infalling dark matter virializes through a reflection from a hard sphere, while in the other it accretes onto a central sink. We find that in both cases, radiative cooling by H_2 affects the collapse dynamics of the gas only after it has already virialized and become part of the bound object. Therefore, radiative cooling is unlikely to have triggered the initial collapse of perturbations at redshifts $z > 10$. Nevertheless,

[1]Astronomy Department, Harvard University, 60 Garden Street, Cambridge, MA 02138, USA
[2]Institute for Advanced Study, Olden Lane, Princeton, NJ 08540, USA
[3]Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 0HA, England

baryonic objects with masses well below the linear-theory Jeans mass ($\lesssim 10^3 M_\odot$) form at high redshifts due to shell crossing by the dark matter. Such objects could be the progenitors of a primordial population of high-mass stars in the intergalactic medium.

2.1 Introduction

Observations indicate that by a redshift ~ 5 the universe contained nonlinear objects such as quasars and absorption systems (see, e.g. review by Woltjer 1990). As current quasar searches are limited from probing even higher redshifts by practical considerations, there is no reason to suspect that they have already revealed the first generation of baryonic objects in the universe. In fact, popular Cold Dark Matter (CDM) cosmologies predict the collapse of the first baryonic objects as early as $z \sim 30$ on a mass-scale just above the cosmological Jeans mass $\sim 10^5 M_\odot$ (Peebles 1993).

The earliest baryonic objects are interesting for a variety of reasons. First, in the bottom-up hierarchical picture of structure formation, they provide the elementary building blocks for larger mass objects that form later. For example, Peebles & Dicke (1968) have suggested that globular clusters may be the fossils of Jeans mass objects that collapsed and fragmented into stars at high redshifts, long before galactic halos formed. Second, direct detection of low-mass objects at $z \gtrsim 10$ can be used to place constraints on the power-spectrum of primordial density fluctuations on very small scales that cannot be probed by microwave background anisotropy experiments (e.g., White et al. 1994). Third, the first baryonic objects may have fragmented into stars and contaminated the intergalactic medium with metals through supernova-driven winds. Recent spectroscopic observations with the Keck telescope indicate the existence of nonzero metalicity even in Lyα clouds (Fan 1994; Tytler & Fan 1995).

The evolution of the first baryonic objects depends on their molecular composition and thermodynamic state. The chemistry and radiative cooling during the cosmological collapse of these objects define a well-posed problem for investigation, since the element abundances are fixed by primordial nucleosynthesis and the cooling rates are not complicated (at least initially) by energy input or metal enrichment from stars. Under primordial conditions, molecular hydrogen (H_2) could become the most efficient coolant of collapsing clouds. Indeed, it has been suggested by various authors (e.g. Saslaw & Zipoy 1967, Peebles & Dicke 1968, Hirasawa 1969, Matsuda et al. 1969, Hutchins 1976, and more recently Palla et al. 1983, Lepp & Shull 1984, and Lahav 1986) that radiative H_2 cooling may, in fact, control the formation of the first generation of objects. However, these suggestions were based on highly simplified models involving homogeneous clouds that cool radiatively while undergoing a uniform free-fall collapse. In reality, a collapsing system becomes highly inhomogeneous as its central density increases. The resulting density profile tends to steepen until a shock develops (see, e.g. Bertschinger 1985 for self-similar solutions well above the Jeans mass). As the shock thermalizes the kinetic energy of the infalling gas, it alters considerably the thermal history of the gas relative to the predictions of the homogeneous collapse models.

In order to study the effect of molecular cooling on collapsing primordial clouds, it is necessary to follow both the dynamics of the gas as a whole and the chemical evolution of its various molecular components. Unfortunately, the full implementation of all relevant chemical reactions and cooling rates in three-dimensional hydrodynamic simulations (Cen & Ostriker 1994; Hernquist, Katz, & Weinberg 1995) requires an extensive computational effort and has not yet been accomplished. As the first step towards a future implementation of this type, we examine in this paper spherically symmetric systems. We use a one-dimensional Lagrangian code to follow the dynamical, thermal and chemical evolution of a spherical system made of collisionless dark matter and collisional gas.

The minimum baryonic mass of the first generation of objects is defined by the smallest scale over which the repulsive pressure force is unable to counteract the attractive gravitational force. Since molecular hydrogen formation becomes more effective as the gas density increases, H_2 cooling can provide a positive feedback that reduces the gas pressure as the density increases, thus allowing the density to increase even further. In this paper we address the question: *could molecular cooling lower the mass of the first baryonic objects down to low (possibly stellar) values?*

The outline of the paper is as follows. In § 2 we review the notion of a cosmological Jeans mass based on linear perturbation theory, and describe how the presence of dark matter, nonlinearities, or cooling, affect this notion. In § 3 we describe the numerical code and its ability to reproduce various analytic results. In § 4 we present numerical results for the evolution of clouds with different initial conditions. In particular, we identify the regime of mass-scales and density contrasts over which radiative cooling by H_2 affects the collapse dynamics. Finally, § 5 summarizes the implications of these results for the first generation of bound objects in the universe.

2.2 Spherical Collapse and the Cosmological Jeans Mass

The Jeans length λ_J was originally defined (Jeans 1928) in Newtonian gravity as the critical wavelength that separates oscillatory and exponentially growing sinusoidal density perturbations in an infinite, uniform, and stationary distribution of gas. On scales smaller than λ_J, the sound crossing time is shorter than the gravitational free-fall time, allowing the build-up of a pressure force to balance gravity. The Jeans mass is defined as the mass within a sphere of radius $\lambda_J/2$, $M_J = (4\pi/3)\rho(\lambda_J/2)^3$. In a perturbation with a mass greater than M_J, self-gravity cannot be supported by pressure, and so the gas is unstable to gravitational collapse. The Newtonian derivation of the Jeans instability suffers from a conceptual inconsistency, as the unperturbed gravitational force of the uniform background must induce bulk motions (cf. Binney & Tremaine 1987). However, this inconsistency is remedied when the analysis is done in an expanding universe.

The perturbative derivation of the Jeans instability criterion can be carried out in a cosmological setting by considering a sinusoidal perturbation superposed on a uniformly expanding background. Here, as in the Newtonian limit, there is a critical wavelength λ_J that separates oscillatory and growing modes. Although the expansion of the background slows down the exponential growth of the amplitude to a power-law growth, the fundamental concept of a minimum mass that can collapse at any given time remains the same (see, e.g. Kolb & Turner 1990, Peebles 1993).

To be specific, let us consider an Einstein-de Sitter universe with $\Omega_{\rm dm}+\Omega_{\rm b} = 1$, where $\Omega_{\rm dm} = \bar{\rho}_{\rm dm}/\rho_{\rm c}$, $\Omega_{\rm b} = \bar{\rho}_{\rm b}/\rho_{\rm c}$, $\bar{\rho}_{\rm dm}$ is the average dark matter density, $\bar{\rho}_{\rm b}$ is the average baryonic density, and $\rho_{\rm c}$ is the critical density. We assume spatial fluctuations in the gas and dark matter densities in the form of a single spherical Fourier mode,

$$\frac{\rho_{\rm dm}(r,t)-\bar{\rho}_{\rm dm}(t)}{\bar{\rho}_{\rm dm}(t)} = \delta_{\rm dm}(t)\frac{\sin(kr)}{kr} \tag{2.1}$$

$$\frac{\rho_{\rm b}(r,t)-\bar{\rho}_{\rm b}(t)}{\bar{\rho}_{\rm b}(t)} = \delta_{\rm b}(t)\frac{\sin(kr)}{kr}, \tag{2.2}$$

where $\bar{\rho}_{\rm dm}(t)$ and $\bar{\rho}_{\rm b}(t)$ are the background densities of the dark matter and baryons, $\delta_{\rm dm}(t)$ and $\delta_{\rm b}(t)$ are the dark matter and baryon overdensity amplitudes, r is the comoving radial coordinate, and k is the comoving perturbation wavenumber. We assume a γ=5/3 ideal gas equation-of-state for the baryons. Initially, at time $t=t_{\rm i}$, the gas temperature is uniform $T_{\rm b}(r,t_{\rm i})$=$T_{\rm i}$, and the perturbation amplitudes are small $\delta_{\rm dm,i}$, $\delta_{\rm b,i}\ll 1$. We define the region inside the first zero of $\sin(kr)/(kr)$, namely $0<kr<\pi$, as the "cloud".

The evolution of $T_{\rm b}(r,t)$ is determined by the coupling of free electrons to the Cosmic Background Radiation (CBR) through Compton scattering, by the adiabatic expansion of the gas, and by other radiative heating and cooling processes. Depending on the detailed chemical composition of the baryons, $T_{\rm b}(r,t)$ will be somewhere between the CBR temperature $T_\gamma \propto t^{-2/3}$ and the adiabatically-scaled temperature $T_{\rm ad}\propto t^{-4/3}$. In the limit of tight coupling to T_γ, the gas temperature remains uniform. On the other hand, in the adiabatic limit, the temperature develops a gradient according to the relation

$$T_{\rm b}\propto \rho_b^{(\gamma-1)}. \tag{2.3}$$

The evolution of $\delta_{\rm dm}(t)$ in the linear regime is described by the equation (Kolb & Turner 1990),

$$\left(\frac{d^2}{dt^2}+\frac{4}{3t}\frac{d}{dt}\right)\delta_{\rm dm}=\frac{2}{3t^2}\left(\Omega_{\rm b}\delta_{\rm b}+\Omega_{\rm dm}\delta_{\rm dm}\right) \tag{2.4}$$

whereas the evolution of $\delta_{\rm b}(t)$ is described by

$$\left(\frac{d^2}{dt^2}+\frac{4}{3t}\frac{d}{dt}\right)\delta_{\rm b}=\frac{2}{3t^2}\left(\Omega_{\rm b}\delta_{\rm b}+\Omega_{\rm dm}\delta_{\rm dm}\right)-\frac{k_{\rm B}T_{\rm i}}{\mu m_p}k^2\left(\frac{t_{\rm i}}{t}\right)^{(8+2\beta)/3}\left(\delta_{\rm b}+\frac{2}{3}\beta[\delta_{\rm b}-\delta_{\rm b,i}]\right). \tag{2.5}$$

Here, m_p is the proton mass, μ is the mean molecular weight of the gas, and $k_{\rm B}$ is the Boltzmann constant. The parameter β distinguishes between the two limits for the evolution of the gas temperature. In the adiabatic limit $\beta=1$, and when the baryon temperature is uniform and locked to the background radiation, $\beta=0$. The last term on the right hand side (in square brackets) takes into account the extra pressure gradient force in $\nabla(\rho_{\rm b}T)=(T\nabla\rho_{\rm b}+\rho_{\rm b}\nabla T)$, arising from the temperature gradient which develops according to equation (2.3) in the adiabatic limit. The Jeans wavelength $\lambda_J=2\pi/k_J$ is obtained by setting the right-hand side of equation (2.5) to zero, and solving for the critical wavenumber k_J. As can be seen from equation (2.5), the critical wavelength λ_J (and therefore the mass M_J) is time-dependent [4]. We infer from equation (2.5) that as time proceeds, perturbations with increasingly smaller initial wavelengths stop oscillating and start to grow. To derive the time-evolution

[4] In the limit $\Omega_{\rm b}\ll\Omega_{\rm dm}\approx 1$ and $T_{\rm b}\approx T_\gamma$ (i.e. $\beta=0$ in eq. [2.5]), the growing solution to equation (2.4) gives $\delta_{\rm dm}\propto t^{-2/3}$, and equation (2.5) can be made homogeneous in time. In this limit, i.e. when almost all the mass is in dark matter and the density of free electrons is sufficiently high to lock the gas temperature to the CBR temperature, λ_J becomes approximately time-independent.

of λ_J, we must solve for the time-evolution of the temperature $T_{\rm b}(r,t)$, including the effects of the various cooling and heating mechanisms, and the evolution of the chemical composition of the baryons.

Another, more fundamental, difficulty is to relate the Jeans mass to the mass of collapsed, bound objects. The above analysis is perturbative (eqs. [2.4] and [2.5] are valid only while $\delta_{\rm b}, \delta_{\rm dm} \lesssim 1$) and can only describe the initial phase of the collapse. As $\delta_{\rm b}$ and $\delta_{\rm dm}$ grow and become larger than unity, the density profiles start to evolve. In particular, since material at various radii collapses at different times, there is no reason to assume that the collapsed object includes the entire mass within the original density peak. Indeed, it has been realized (Larson 1969) that in the non-linear regime the density and temperature profiles steepen considerably and the central part of the sphere becomes exceedingly dense and hot. An object may form in the center, and subsequently accrete material from the original peak. In the limit of high masses ($\lambda \gg \lambda_{\rm Jeans}$) and in the absence of radiative cooling, the gas particles virialize through a shock that propagates outwards from the center of the sphere. The shock structure and the evolution of the density profile is self-similar, and near the center the density approaches the power-law profile $\rho \propto r^{-2.25}$ (this was first predicted by Gott 1975, and later proved via a semi-analytic approach by Bertschinger 1985). Although the initial perturbation has a definite mass, it is not clear what mass to associate with the end-product of the collapse. The relation between the mass of the bound object $M_{\rm obj}$ and the parameters of the initial perturbation can be expressed as $M_{\rm obj} = M_{\rm obj}(k, \delta_{\rm dm,i}, \delta_{\rm b,i}, t_{\rm i})$. This relation cannot be inferred directly from linear theory.

In the presence of radiative cooling, the dynamical evolution of the gas departs from self-similarity. In the limit of very efficient cooling, gas particles follow free-fall trajectories until they come arbitrarily close to the origin. If the cooling time is comparable to the dynamical time, the gas particles go through various phases of partial pressure support (see, e.g., Thoul & Weinberg 1995). In general, radiative cooling will facilitate the collapse of the clouds, since it reduces the pressure support of the gas. Thus we expect that $M_{\rm obj}(k, \delta_{\rm dm,i}, \delta_{\rm b,i}, t_{\rm i})$ will be increased relative to its value in the absence of cooling.

The Jeans mass derived from linear theory specifies whether an initial perturbation, characterized by the parameters k, $\delta_{\rm dm,i}$, $\delta_{\rm b,i}$ and $t_{\rm i}$, starts to grow. To determine the mass of the bound object resulting from its collapse we must follow numerically the nonlinear evolution of the density peak, including all microphysical gas processes. In the following sections, we address this problem using a one-dimensional spherically symmetric hydrodynamical code.

2.3 The numerical code

3.1. *Description of the code*

The code we use is a modified version of the one-dimensional, spherically symmetric Lagrangian hydrodynamics code written by Thoul (Thoul & Weinberg 1995). The code evolves a mixture of two fluids by moving concentric spherical shells of fixed mass in the radial direction. A description of the numerical method and the details of the original code can be found in the above reference. Here we only summarize the equations solved by the code, and describe briefly the new additions: chemical reactions, Compton cooling, and cooling by the H_2 molecule.

The Lagrangian equations of hydrodynamics in spherical symmetry are the continuity equation for the baryons:

$$\frac{dm_{\rm b}}{dr_{\rm b}} = 4\pi r_{\rm b}^2 \rho_{\rm b}, \tag{2.6}$$

the momentum equations for the baryons and dark matter:

$$\frac{dv_{\rm b}}{dt} = -4\pi r_{\rm b}^2 \frac{dp}{dm_{\rm b}} - \frac{M(r_{\rm b})}{r_{\rm b}^2}, \tag{2.7}$$

$$\frac{dv_{\rm dm}}{dt} = -\frac{M(r_{\rm dm})}{r_{\rm dm}^2}, \tag{2.8}$$

the energy equation:

$$\frac{du}{dt} = -p\frac{d(\rho_{\rm b}^{-1})}{dt} + \frac{\Lambda}{\rho_{\rm b}}, \tag{2.9}$$

and the equation of state:

$$p = (\gamma - 1)\rho_{\rm b} u. \tag{2.10}$$

Here $\gamma = 5/3$ is the adiabatic index, and we have set $G = 1$. Each baryonic shell is described by its radius $r_{\rm b}$, velocity $v_{\rm b}$, mass density $\rho_{\rm b}$, pressure p, internal energy per unit mass u, and the baryonic mass interior to it $m_{\rm b}$. The dark matter shells are described by their radius $r_{\rm dm}$, velocity $v_{\rm dm}$, and the enclosed dark matter mass $m_{\rm dm}$. $M(r)$ is the total (baryonic plus dark matter) mass inside a shell at radius r. The details of the cooling terms included in Λ in equation (2.9) are given in Appendix A.

The baryonic component is composed of the nine species H, H^-, H^+, He, He^+, He^{++}, H_2, H_2^+, and e^-. For the $i^{\rm th}$ species, we denote the total number of particles by N_i, the number density and molecular weight of these particles by n_i and μ_i respectively, and the mass fraction by

$$y_i = \frac{\mu_i n_i}{\sum_{j=1}^9 \mu_j n_j} = \frac{\mu_i n_i}{\rho_{\rm b}}. \tag{2.11}$$

The relative abundances of helium species ($i = 4, 5, 6$) are denoted by $x_i = n_i/n_{\rm He,tot}$, and the relative abundance of the hydrogen species ($i = 1, 2, 3, 7, 8$) and electrons ($i = 9$) are given by $x_i = n_i/n_{\rm H,tot}$, where $n_{\rm He,tot} = (n_{\rm He} + n_{\rm He^+} + n_{\rm He^{++}})$ and $n_{\rm H,tot} \equiv (n_{\rm H} + n_{\rm H^+} + n_{\rm H^-} + 2n_{\rm H_2} + 2n_{\rm H_2^+})$.

The above nine species interact through various chemical reactions. Only the global quantity $\rho_{\rm b}$ appears in the hydrodynamical equations (2.6)-(2.10). After each shell has been moved to its new location by a hydrodynamical iteration, the nine species are allowed to interact for a time $dt_{\rm H}$, and the relative abundances in each shell are upgraded according to the rate equations

$$\frac{dn_i}{dt} = \sum_{l=1}^{9}\sum_{m=1}^{9} a_{lmi} k_{lm} n_l n_m + \sum_{j=1}^{9} b_{ji} k_j n_j. \tag{2.12}$$

Here the k_{lm} denote reaction rate coefficients, k_n represent photo-ionization or photo-dissociation integrals, and $a_{lmi}, b_{ni} = 0, \pm 1$, or ± 2, depending on the reaction. The next hydrodynamical iteration starts at a time interval $dt_{\rm H}$ later, with the updated relative abundances. In Appendix B we provide a table of the chemical reactions, and the adopted rate coefficients and cross sections.

A complication to the above numerical approach arises due to the stiffness of the system of differential equations (2.12). The stiffness increases the required computation time by a large factor, because standard solver routines for differential equations require exceedingly small time-steps, while routines designed to solve a stiff system of equations require many matrix operations with 9×9 matrices, in our case. We have experimented with each routine in *Numerical Recipes* (Press et al. 1992), and have found the routine *stifbs* to be the most efficient. We also compared the performance of *stifbs* with the more popular routine LSODAR (Hindmarsh 1983) to solve the system (2.12). We found *stifbs* to produce identical answers $\sim 20\%$ faster than LSODAR. The nine equations in system (2.12) could be reduced to six, using the conservation equations for the hydrogen and helium mass and for the charge. However, we found that this reduction actually *increases* the computation time, because the use of the algebraic conservation laws requires that the calculation of each number density n_i be done to a much higher degree of accuracy than otherwise necessary. Another simplification could arise if the species were either "frozen", or reached equilibrium within a hydrodynamical iteration time. Unfortunately, equations (2.12) do not have a single time constant and the concentrations of different species change on different time-scales; in particular, $n_{\rm H_2}$ evolves on the hydrodynamic time-scale. The only simplification we were able to make was to freeze the helium abundances during the initial phase of the dynamical evolution, thus reducing the number of equations to be solved from 9 to 6.

We adopt two different central boundary conditions, corresponding to placing either a hard sphere or a sink (black hole) in the center (see Thoul and Weinberg 1995). In the first case, the dark matter virializes by bouncing off the hard sphere, while in the latter case the dark matter falls freely into a central sink. The first case may provide a more realistic representation of collapsing regions in the universe, for which non-spherical motions (or angular momentum) define an inner boundary to the infall. Quantities for the n^{th} shell are always calculated using information from the n^{th} and $(n+1)^{th}$ shell, thus boundary effects from the surface of the sphere propagate inwards by one shell at each iteration. In order to avoid unphysical contamination of the results by signals propagating from the outermost shell in the simulation, we choose the outer boundary condition to be such that the total perturbed mass inside the outermost shell is zero, *and* the density profile has zero slope at the surface. Both of these conditions are satisfied at the first minimum of the $\sin(x)/x$ function in equations (2.1)-(2.2), i.e. at x=4.4934. This choice eliminates both pressure and gravitational forces on the outermost shell, and allows this shell to simply follow the general expansion of the universe.

3.2 *Tests of the code*

Although Thoul & Weinberg (1995) have made extensive tests of the code, we performed a few additional tests. The most important side-effect of the addition of the chemical rate equations was an increase in computation time by approximately a factor of 10. This forced us to use a lower resolution (500 gas shells and 5000 dark matter shells) in our runs. However, comparisons with runs of five times higher resolution (2500 gas shells, 25000 dark matter shells) showed no considerable differences in the trajectories of the shells.

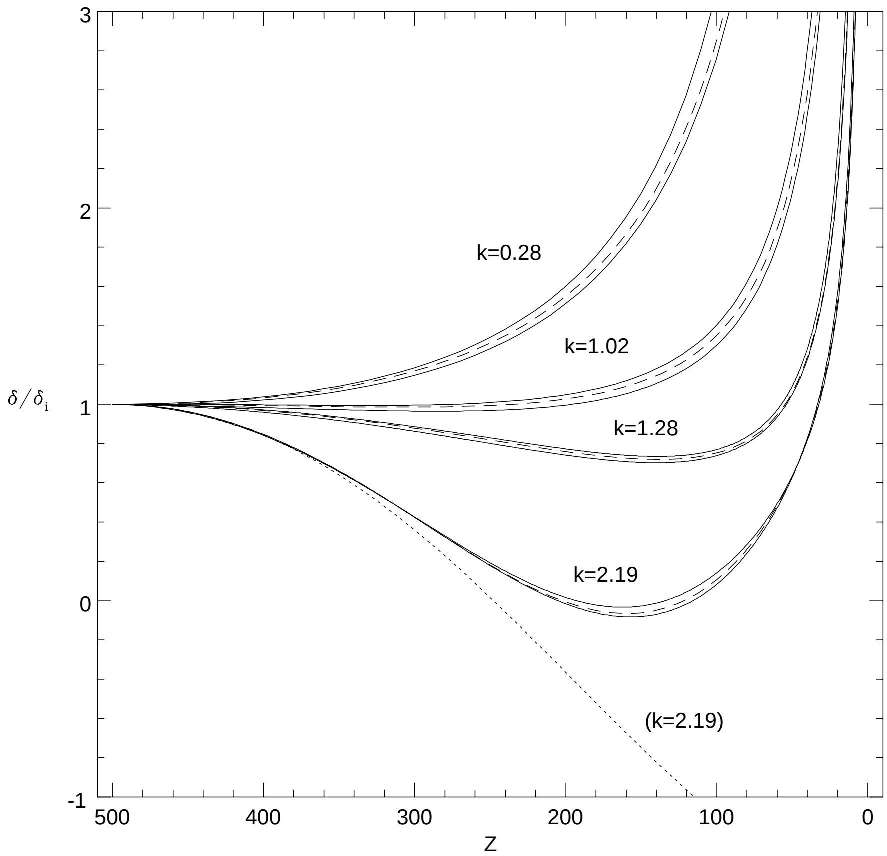


Figure 2.1: *Evolution of the overdensity for single spherical Fourier modes in the linear regime. The overdensity δ is normalized to its initial value δ_i. The comoving wavenumber k is given in* kpc^{-1}. *The dashed lines are obtained analytically from linear theory. The pairs of solid lines show numerical results for $\delta(r, z)$ at the center and at the surface of the sphere. The overdensity at the center of the sphere becomes slightly higher than at the surface. The dotted line shows the analytic results from linear theory when the temperature gradient term in the square brackets in equation (2.5) of § 2 is ignored.*

The numerical code contains three almost independent modules. These modules perform the hydrodynamical, chemical, and cooling calculations. We tested these three modules as follows.

(a) Hydrodynamics

The hydrodynamics part of the code was tested by Thoul & Weinberg (1995). They reproduced analytical solutions for pulsations of a polytrope, and for the self-similar collapse of a spherically symmetric, cosmological perturbation (Bertschinger 1985). In order to test the code in a context closer to the present application, we checked that it reproduces the evolution of the density contrast in the linear regime. The linear theory for the evolution of the density contrast of a single spherical Fourier mode was described in § 2 above. Supplemented with the temperature evolution $T_{\rm b}(t)$, this calculation yields the factor $\delta_k(t)$ in

$$\delta(r,t) \equiv \frac{\delta\rho}{\rho}(r,t) = \delta_k(t)\frac{\sin(kr)}{kr} \tag{2.13}$$

for each value of k for both the dark matter and the gas. We ran the code with an initial overdensity profile given in equations (2.1)-(2.2) for various values of k and $\delta_k(t=0) = 0.05$. We found that the shape of the profile stays unchanged to a high precision, while the overdensity changes by an overall factor very close to the value predicted from linear theory. Figure 2.1 shows how closely the overdensity in three gas shells follow the predicted $\delta_k(t)$, for four different values of k.

(b) Chemistry

To test the numerical integrations of the system of chemical reactions (2.12), we checked that the total hydrogen and helium mass and the total charge are conserved, or equivalently that the quantities

$$N_{\mathrm{H,tot}} = N_{\mathrm{H}} + N_{\mathrm{H^+}} + N_{\mathrm{H^-}} + 2N_{\mathrm{H_2}} + 2N_{\mathrm{H_2^+}} \tag{2.14}$$

$$N_{\mathrm{He,tot}} = N_{\mathrm{He}} + N_{\mathrm{He^+}} + N_{\mathrm{He^{++}}} \tag{2.15}$$

$$Q_{\mathrm{tot}} = N_{\mathrm{H^+}} - N_{\mathrm{H^-}} + N_{\mathrm{He^+}} + 2N_{\mathrm{He^{++}}} + N_{\mathrm{H_2^+}} - N_{\mathrm{e}}, \tag{2.16}$$

are constant in time in each shell. We found all three quantities to be conserved to at least one part in 10^8. We also checked that the abundances converge to the correct equilibrium values. We ran the routine *stifbs* starting with the set of initial concentrations $x_{\mathrm{H_2}} = 1.7 \times 10^{-6}$, $y_{\mathrm{He}} = 0.24$, and $y_{\mathrm{H}} \approx 0.76$ at various temperatures, until none of the abundances changed by more than 0.1% from one step to the next. The equilibrium abundances of $\mathrm{H_2}, \mathrm{H^-}$, and $\mathrm{H_2^+}$ are always very small, so by eliminating these 3 species from equations (2.12), and by setting $d/dt = 0$, one can find algebraically the expected equilibrium abundances of the six remaining species. We found that *stifbs* indeed converged to these values. Figure 2.2 shows the equilibrium abundances of all nine species vs. temperature.

(c) Cooling and Chemistry

Several authors have considered molecular cooling of a uniform sphere in free-fall collapse (e.g. Palla et al. 1983, Lepp & Shull 1984). The general conclusion from these studies can be summarized as follows. On a temperature vs. density plot, the temperature departs from the adiabatic relation, $T_{\mathrm{b}} \propto \rho_{\mathrm{b}}^{2/3}$, due to $\mathrm{H_2}$ cooling and achieves a local maximum T_{max}, after which it turns around and declines as the density increases further. The temperature does not start to rise again over several decades of increase in density. The exact values of T_{max}, the density at the turnaround in temperature, and the density at which the temperature starts to rise again, all depend on the adopted reaction rates and cooling function, as well as on the initial size, density, temperature and composition of the sphere.

We therefore tested only the general conclusion that a turnaround in temperature should occur and that the temperature should not increase over the next several decades in density. We considered a uniform sphere in a homogeneous free-fall collapse with the following initial parameters: $R = 10^{21}$cm, $\rho_0 = 4.52 \times 10^{-26}$g cm^{-3}, $T = 64$K, $x_{\mathrm{H_2}} = 1.7 \times 10^{-6}$, $y_{\mathrm{He}} = 0.24$, and $y_{\mathrm{H}} \approx 0.76$. The chemical reactions and cooling mechanisms are summarized in Appendices A and B. The resulting temperature vs. density plot, as well as the relative abundance of $\mathrm{H_2}$ vs. density, are shown

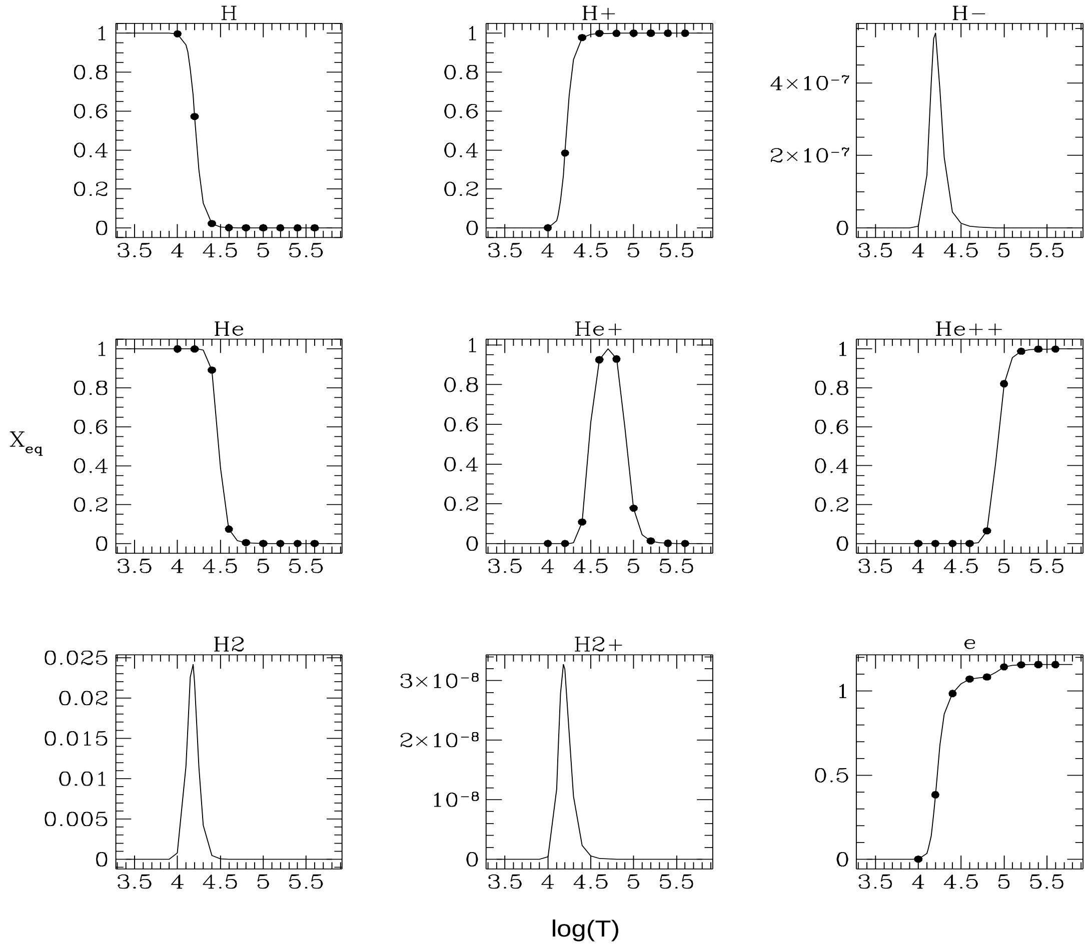


Figure 2.2: *Equilibrium abundances of the nine species as a function of temperature. The abundances were obtained by integrating equations (2.12) with the routine* stifbs *until the abundances reached their equilibrium values. The dots show values obtained algebraically from equations (2.12) by setting n_{H_2}=n_{H^-}=$n_{H_2^+}$=0 and d/dt=0.*

in Figure 2.3. As seen in this figure, turnaround indeed occurs at $\rho/\rho_0 = 80$ when the temperature reaches the maximum value $T_{\max} \sim 400$K. We note that the relative abundance of H_2 reaches $x_{H_2} = 3.8 \times 10^{-3}$, which is consistent with the H_2 fractions in Palla et al. (1983), and Lepp & Shull (1984).

2.4 Numerical Results and Discussion

2.4.1 Initial Conditions

For all cases presented in this paper, we adopt a Hubble constant of $H_0 = 50$ km s^{-1} Mpc^{-1} ($h = 0.5$). In runs with dark matter we assume an Einstein-de Sitter universe with $\Omega_{\rm dm} = 0.9$, $\Omega_{\rm b} = 0.1$, and in runs without dark matter, we assume an open universe with $\Omega_{\rm b} = 0.1$. We start the integrations at $z_i = 500$, when the perturbations are still in the linear regime. The temperature for the CBR is taken to be $T_\gamma = 2.726(1 + z)\,$K (Mather et al. 1994), and the initial temperature of the

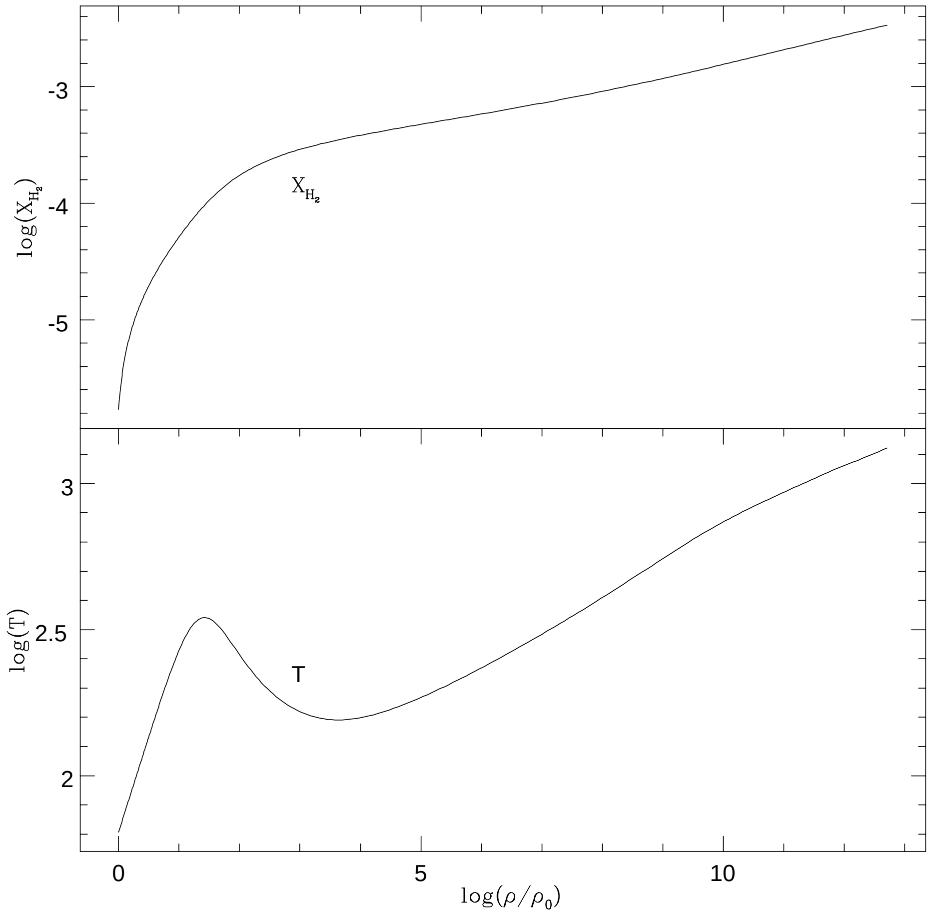


Figure 2.3: *Temperature (in K) and fractional abundance of* H_2 *molecules vs. density in a freely collapsing uniform sphere in the presence of* H_2 *cooling. Parameters of the sphere are described in § 3.2(c).*

cloud is assumed to be uniform and equal to T_γ at z_i. We use spherical k-modes as the initial overdensity profiles,

$$\rho_{\mathrm{dm,i}}(r) = \Omega_{\mathrm{dm}}\rho_{\mathrm{c}}[1+\delta_{\mathrm{dm,i}}\frac{\sin(kr)}{kr}] \tag{2.17}$$

$$\rho_{\mathrm{b,i}}(r) = \Omega_{\mathrm{b}}\rho_{\mathrm{c}}[1+\delta_{\mathrm{b,i}}\frac{\sin(kr)}{kr}], \tag{2.18}$$

where $\rho_{\mathrm{c}} = (6\pi G t^2)^{-1}$ is the critical density. With this choice, the evolution in the linear regime can be compared directly to the prediction of equations (2.4)-(2.5). As mentioned at the end of § 3.1, this choice also eliminates numerical edge effects at the outer boundary. We have experimented with alternative overdensity profiles, such as a Gaussian, or the most probable profile for random Gaussian fluctuations (Bardeen et al. 1986, hereafter BBKS), and found that our qualitative conclusions are insensitive to the detailed shape of the initial density profile.

The density profiles in equations (2.17)-(2.18) can be used to define three characteristic mass-scales. First, the total baryonic mass which is gravitationally bound includes all the baryonic material enclosed by the radius corresponding to the first *minimum* of $\sin(kr)/(kr)$, i.e. inside $kR_{\mathrm{bound}} = 4.4934$. We denote this mass as

Table 2.1: *The r.m.s. fluctuation $\sigma(M)$ in two different filters with the COBE-normalized ($\sigma_8 = 1.22$) BBKS power-spectrum of a cold dark matter universe at z=500.*

$M_{\rm cloud}$	$\sigma_{\rm dm}(M, z=500)$	
$(M_\odot)$	Top-hat	Gaussian
50	0.118	0.113
100	0.115	0.110
500	0.107	0.103
1,000	0.104	0.100
5,000	0.097	0.093
10^4	0.093	0.090

$M_{\rm bound}$. Second, we define the mass of the baryonic material which has a positive overdensity (i.e. contained within the first *zero* of $\sin(kr)/(kr)$, at $kR_{\rm cloud} = \pi$) to be $M_{\rm cloud}$. Finally, the baryonic mass which has virialized and condensed into a central object, $M_{\rm obj}$, is redshift-dependent and will be defined more precisely in the next section.

The initial amplitude of the dark matter density profile perturbation can be written as $\delta_{\rm dm,i} = \nu\sigma$, where σ is the *r.m.s.* density fluctuation within some filter of a given shape and size in the CDM model. The value of ν characterizes the likelihood of the occurance of the peak in a Gaussian density field. This characterization depends weakly on the shape and size of the filter, and strongly on the choice of power-spectrum. For reference, we quote in Table 2.1 the value of σ for six different mass-scales $M_{\rm cloud}$, and two different filter shapes with the COBE normalized BBKS power spectrum. The values in this table were obtained under the normalization $\sigma_{8h^{-1}{\rm Mpc}} = 1.22$ (for $\Omega_{\rm dm} = 0.9$, and $h = 0.5$; cf. Bunn et al. 1995), using a filter radius $R_f = 0.12R_{\rm cloud}$. The initial amplitude of the baryon density perturbation is a small fraction of the dark matter amplitude, because the baryons had been locked to the CBR until $z \approx 10^3$, while the dark matter fluctuations have had more time to grow since matter-radiation equality. We chose the somewhat arbitrary value $\delta_{\rm b,i} = \frac{1}{10}\delta_{\rm dm,i}$ to take this effect into account. Because of the tight coupling of the baryons to the CBR before the recombination epoch, the initial velocity of each baryonic shell at $z_i = 500$ is assumed to be the Hubble velocity $v_{\rm b}(r) = Hr$. For the dark matter, we assume that the perturbation is dominated by the growing mode, and choose the initial velocity for the dark matter shells accordingly, $v_{\rm dm}(r) = Hr(1 - \frac{1}{3}\langle\delta_{\rm dm}\rangle_r)$.

The initial relative abundances and mass fractions of the various species are taken to be: $x_{\rm H^+}$=$x_{\rm e}$=10^{-4} (cf. Peebles 1993), $x_{\rm H_2}$=1.7×10^{-6} (cf. Palla et al. 1984), $x_{\rm H^-}$=$x_{\rm He^+}$=$x_{\rm He^{++}}$=$x_{\rm H_2^+}$=0, and ($y_{\rm H}$+$y_{\rm H^+}$+$y_{\rm H_2}$+$y_{\rm e}$)=0.76, $y_{\rm He}$=0.24.

2.4.2 Definitions of "Collapse" and "Object"

In order to express the state of the gas as a function of initial conditions, we need to define what we mean by the concepts of "collapse" and "object". We base our heuristic definitions on the spherical top-hat collapse model. In this model, virialization occurs

Table 2.2: *Parameters of the numerical runs. Initial heights and widths are at* $z = 500$.

Run Number	Baryon Mass $M_{\rm cloud}(\rm M_\odot)$	Height $\delta_{\rm i}$	Width $R_{\rm cloud}(\rm pc)$	Baryons $\Omega_{\rm b}$	Dark Matter $\Omega_{\rm dm}$	Cooling Yes/No
1	5×10^7	0.13	235.56	0.1	0	No
2	5×10^5	0.13	50.75	0.1	0	No
3	5×10^7	0.39	228.96	0.1	0	No
4	5×10^5	0.39	49.33	0.1	0	No
5	500	0.26	5.13	0.1	0.9	Yes
6	$1,000$	0.26	6.46	0.1	0.9	Yes
7	$1,000$	0.39	6.45	0.1	0.9	Yes
8	$5,000$	0.26	11.05	0.1	0.9	Yes

at half the turnaround radius when the overdensity is $\sim 100-200$. The value appears to be rather insensitive to the underlying cosmology (cf. Fig. 4 in Richstone, Loeb & Turner 1992). We therefore define a shell to have *collapsed* if the average overdensity enclosed by that shell is $\bar{\delta}_{\rm b} \geq \bar{\delta}_{\rm thr} = 200$.

As time evolves, the density profile within the collapsing region steepens, and at some definite time, the average density contrast within the innermost baryonic shell reaches the threshold value $\bar{\delta}_{\rm b}(r_1) = \bar{\delta}_{\rm thr} = 200$. The subsequent shells reach this threshold some time later. We define an *object* to include all the baryonic material within the outermost shell that has collapsed. According to this definition, the mass of the object $M_{\rm obj}(z)$ is zero until the first shell reaches $\bar{\delta}_{\rm thr} = 200$, and is monotonically increasing aftwerwards.

Unlike definitions based on optical depth, or absolute density, the threshold overdensity does not strongly depend on the mass-scale involved. In the presence of cooling we find our results to be rather insensitive to the exact value of the threshold overdensity (see below).

2.4.3 Collapse without dark matter or cooling

The simplest system to consider consists of a purely baryonic cloud following adiabatic infall, without any dark matter or radiative processes. We present results from test cases of such a system in order to establish a prototype against which we will compare results later, when the dark matter component and the radiative processes will be added.

For the initial conditions (cf. § 4.1), it is necessary to fix two free parameters: the height of the density peak $\delta_{\rm b,i}$, and the mass-scale $M_{\rm cloud}$ or equivalently the width $\sim k^{-1}$ of the initial density peak. $\delta_{\rm b,i}$ sets the redshift at which the perturbation reaches the nonlinear phase, and k determines the initial ratio of pressure to gravitational forces.

In Table 2.2 we summarize the parameters for eight different runs. The runs

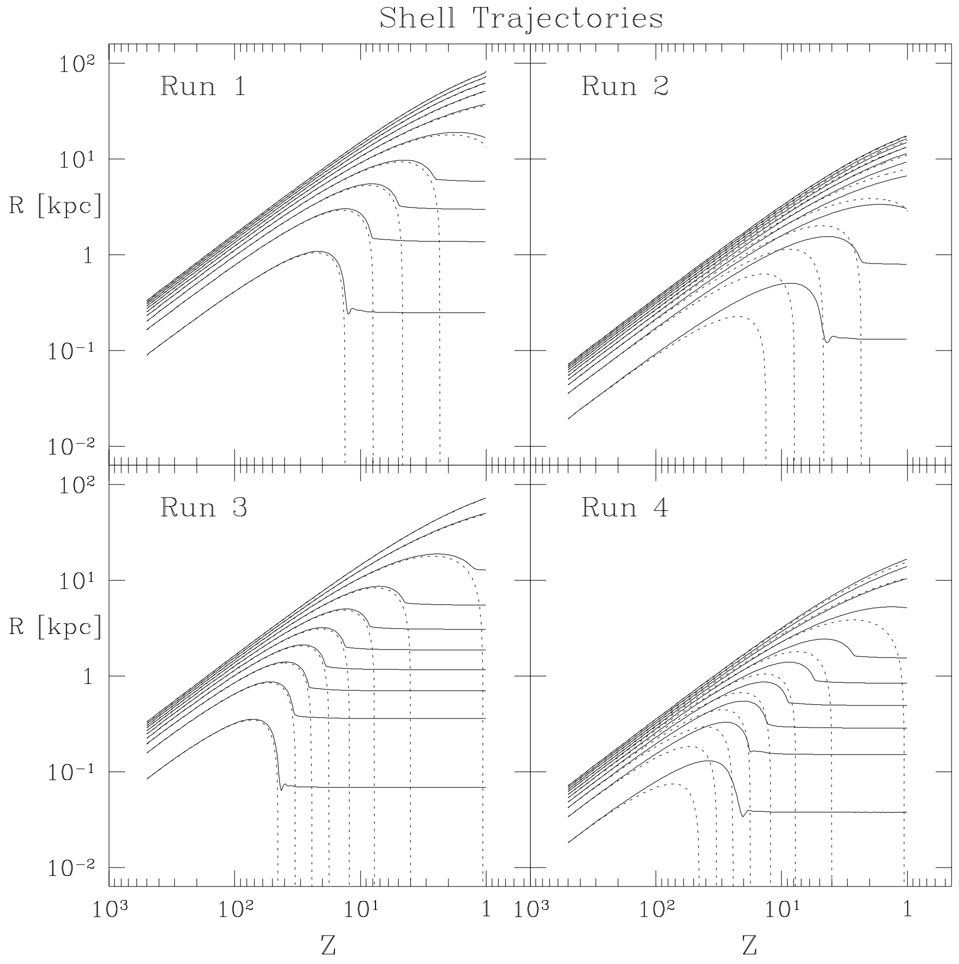


Figure 2.4: *Evolution of the radii of 10 gas shells (enclosing 7, 17, 27, ..., 87, and 97% of the total baryonic mass $M_{\rm bound}$) for four pure baryonic runs without cooling. The parameters of these runs are summarized in Table 2.2. The solid lines show the shell trajectories from the simulations, while the dashed lines show for reference the corresponding shell trajectories for a pressureless gas.*

numbered 1-4 are purely baryonic and do not include any radiative cooling, while the runs numbered 5-8 contain a mixture of baryons and cold dark matter, and include radiative cooling.

We first discuss the results of the first four simulations. In Figure 2.4 we show the trajectories of 10 shells (enclosing 7, 17, 27, ..., 87, and 97% of the total baryonic mass $M_{\rm bound}$) across the radius of the sphere (solid curves), along with the trajectories that each shell would have had if it were freely falling (dashed curves). In all cases, the baryonic shells expand, turn around, collapse, shock, and then drift slowly towards the center. The trajectories of the shells are similar to the semi-analytic solution of Bertschinger (1985). Here, however, the shells hit the shock at slightly different fractions of their turnaround radii rather than at the fixed fraction of $\frac{1}{2}$, due to corrections of order δ_i^2 to Bertschinger's results. The higher mass runs (numbers 1 and 3) initially follow the free-fall trajectories, but eventually deviate from them after the shells shock. In the lower mass runs (numbers 2 and 4) pressure is important and the collapse is delayed relative to the free-fall case even before the

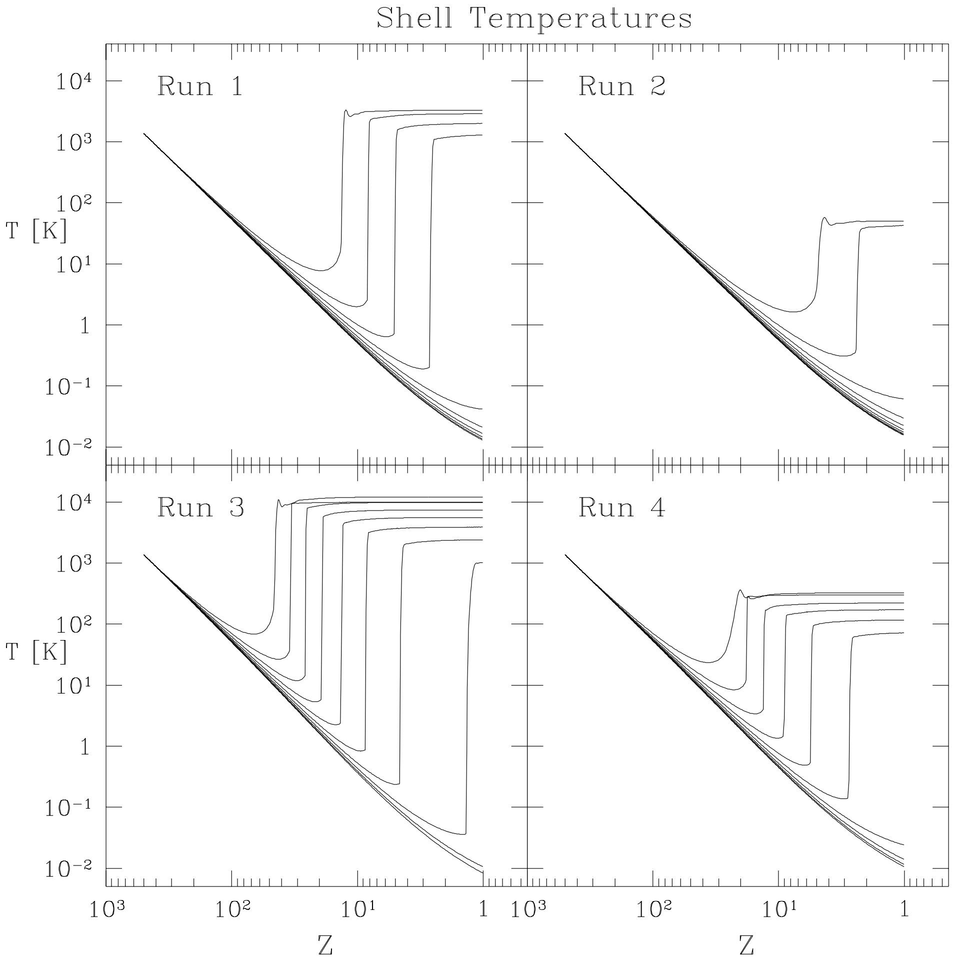


Figure 2.5: *Temperature evolution of the gas shells that appear in Figure 2.4.*

shock. The temperature evolution for these shells is shown in Figure 2.5.

In Figure 2.6 we show the fraction of the total bound baryonic mass that has collapsed to form the object (solid lines). We also show what this collapsed fraction would be for a pressureless freely-falling gas (dotted lines). The higher mass runs deviate from the free-fall curves only after the shocked shells collapse. In the lower mass runs, the collapse is delayed relative to the free-fall case at all times.

2.4.4 Collapse with dark matter and cooling

Given some choice of cosmological parameters H_0, $\Omega_{\rm dm}$ and $\Omega_{\rm b}$, and the ratio $\delta_{\rm b,i}/\delta_{\rm dm,i}$, there are again only two free parameters left, namely the initial baryonic cloud mass $M_{\rm cloud}$, and the amplitude $\delta_{\rm dm,i}$. We have performed simulations on a grid in the $(M_{\rm cloud}, \delta_{\rm dm,i})$ parameter space. The details of the four simulations numbered 5-8 in Table 2.2 are shown in Figures 2.7-2.12.

Figure 2.7 shows the trajectories of 10 shells (enclosing 7, 17, 27, ..., 87, and 97% of the total baryonic mass $M_{\rm bound}$). The trajectory of each shell is shown by a solid line, until the average density contrast within it is smaller than the threshold value $\bar{\delta}_{\rm thr} = 200$. Once the overdensity within a shell reaches $\bar{\delta}_{\rm thr}$, it is considered to

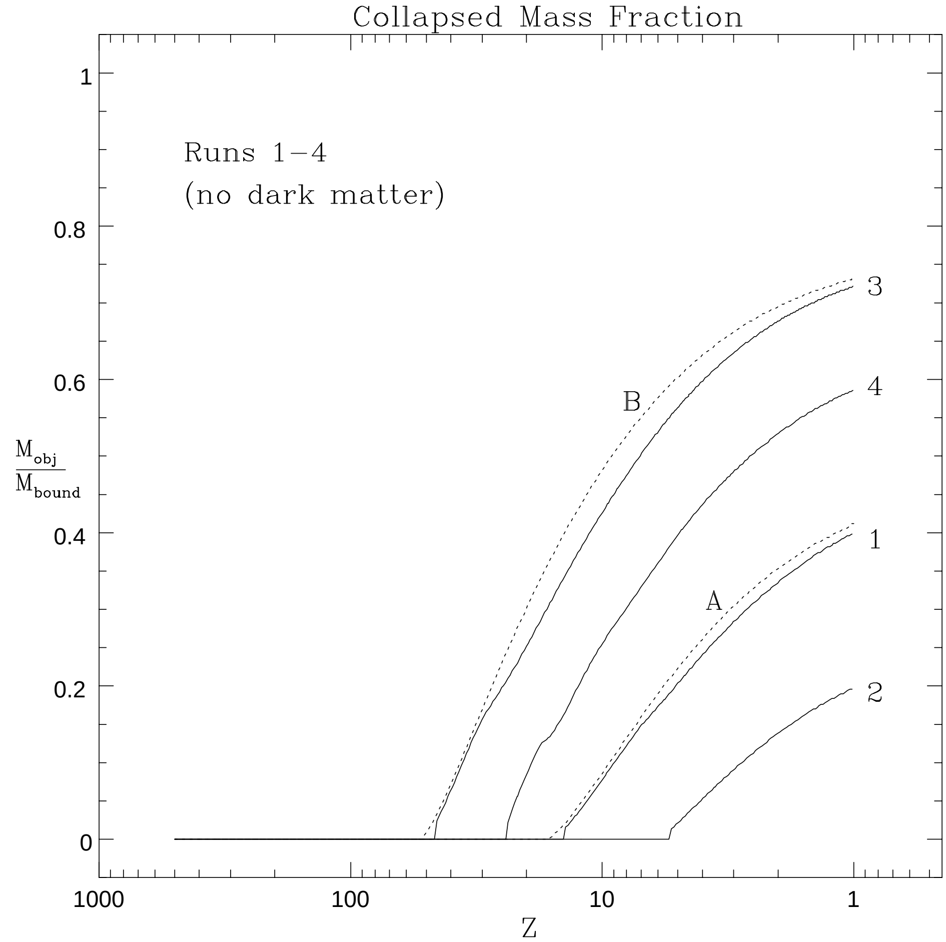


Figure 2.6: *The solid lines show the fraction of the bound mass that has collapsed as a function of redshift for the four runs 1-4 in Table 2.2. The two dashed lines show what the collapsed baryonic mass fraction would be in each case if the cloud had been in free-fall. The line marked "A" is for* $\delta_{\rm dm,i} = 0.13$ *(to be compared with runs 1 and 2) and the line marked "B" is for* $\delta_{\rm dm,i} = 0.39$ *(to be compared with runs 3 and 4).*

be part of the central object. To indicate this in the plots, we continue to draw the shell trajectories as dotted lines after they reach $\bar{\delta} = \bar{\delta}_{\rm thr}$. The dashed lines show the trajectories of the dark matter shells. In the smallest mass run (number 5) cooling is inefficient and the gas shells stop moving after they shock. In larger mass cases (runs 6 and 7) cooling is more efficient. Finally, in the largest mass case (run 8) all the shocked shells cool rapidly and fall towards the center.

Figure 2.8 shows the evolution of the temperature with redshift for the same gas shells. The dotted lines show the CBR temperature $T_\gamma \propto (1+z)$, and the dashed lines show the adiabatic temperature scaling in the linear regime $T_{\rm ad} \propto (1+z)^2$. The temperature of the gas remains approximately uniform initially, with a value in-between the adiabatic and CBR cases. After turnaround the temperature increases adiabatically and a temperature gradient develops. Eventually, as each shell encounters the shock, its temperature jumps abruptly to the virial temperature. In the absence of cooling, the temperature of each shell remains approximately constant, increasing slowly due to the slight adiabatic compression. When cooling is efficient, i.e., when the cooling time becomes smaller than the dynamical time, the temperature

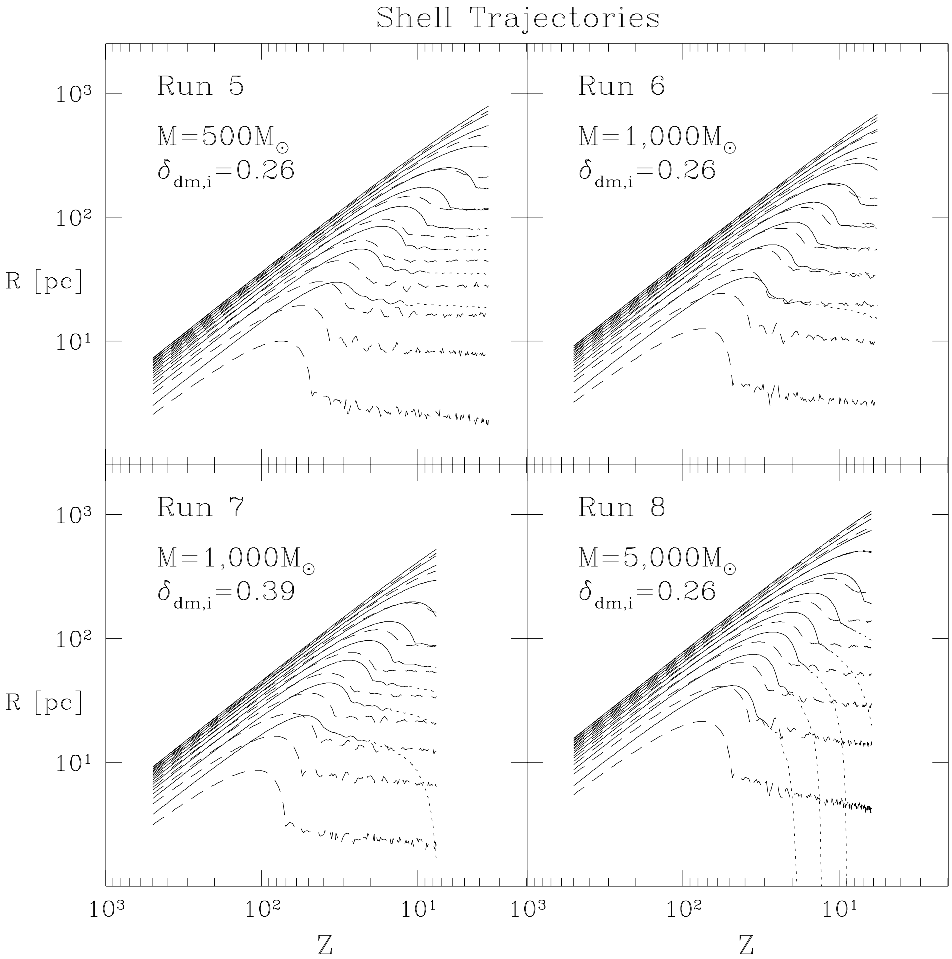


Figure 2.7: *Evolution of the radii of 10 gas shells (enclosing 7, 17, 27, ..., 87, and 97% of the total baryonic mass $M_{\rm bound}$) for four mixed baryons and dark matter runs with cooling. The parameters of these runs, numbered 5-8, are summarized in Table 2.2. The trajectories of dark matter shells are shown as dashed lines. The trajectories of the gas shells are shown as solid lines until they reach $\bar{\delta} = \bar{\delta}_{\rm thr}$, and as dotted lines afterwards.*

reaches a maximum value and then decreases very quickly. The temperature at which this turnover occurs in our simulations is of the order $\log(T/{\rm K}) \approx 2.3-2.8$ at the onset of H_2 cooling. Figure 2.9 shows the evolution of the four most efficient cooling and heating terms with redshift in a typical run. The most efficient radiative mechanism before the shock (when the temperature of the gas is below the temperature of the CBR) is Compton heating; this process keeps the gas temperature above $T_{\rm ad}$. After the shock, the most efficient radiative mechanism is H_2 cooling, causing a turnover in the temperature. Note that the gas temperature reaches a value above T_γ, so that the H_2 molecules as well as Compton scattering cool the gas, rather than heat it. After the temperature starts to decline, the shells fall towards the center essentially at the free-fall rate. When the shells reach a very small radius and the numerical timestep necessary to resolve their motion becomes too small, we artificially set the temperature of these shells to $T = 200$ K and subsequently ignore them.

Figure 2.10 shows the evolution of the relative abundances of H_2 and e^- within five different shells (enclosing 2, 7, 12, 17, and 22% of the total baryonic mass $M_{\rm bound}$).

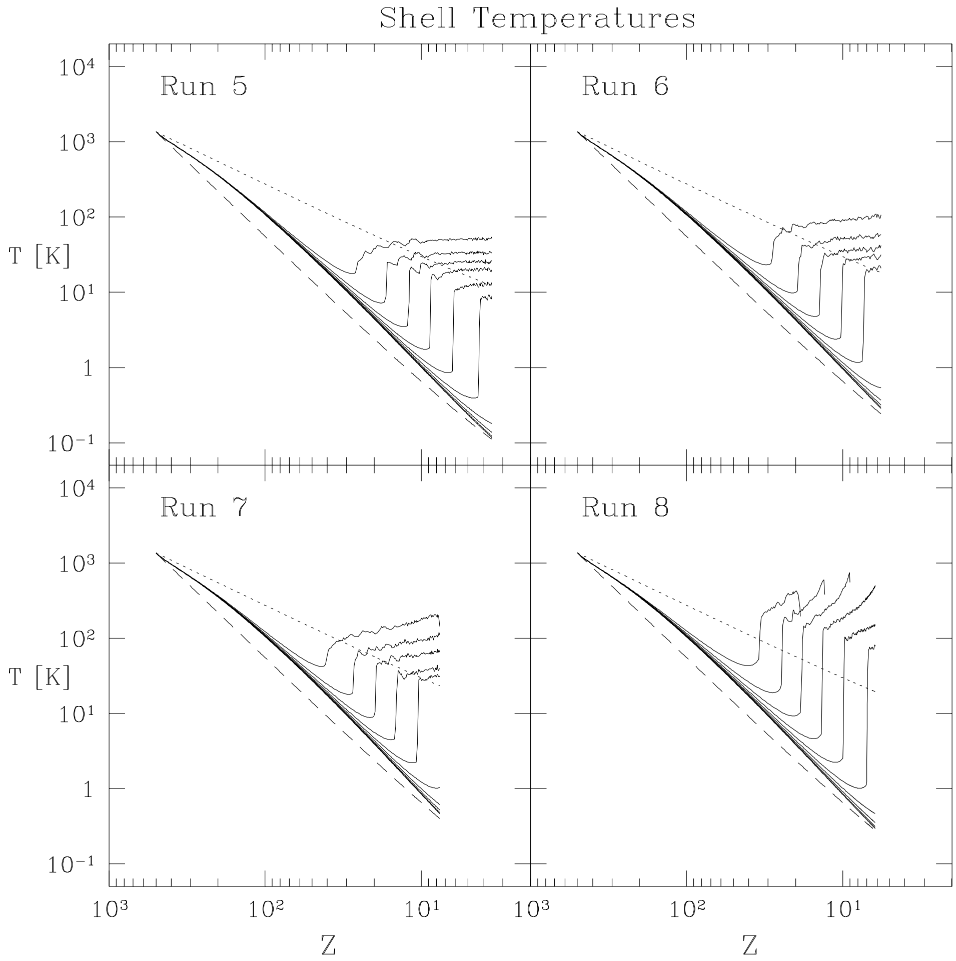


Figure 2.8: *Temperature evolution of the gas shells that appear in Figure 7. The dotted lines show the CBR temperature $T_\gamma \propto (1+z)$, and the dashed lines show the adiabatic temperature $T_{\rm ad} \propto (1+z)^2$. The gas temperature remains between T_γ and $T_{\rm ad}$ until it shocks.*

The highest relative abundance of H_2 is achieved in the case of the largest perturbed mass (run 8), where it reaches about 10^{-3}. The abundances develop a gradient across the sphere, which is positive for H_2 and negative for the electrons. The evolution of the abundances are relatively universal and look similar in all four runs. In particular, $\log x_{H_2} \approx -3.6$ in all four runs before the gradient develops, and the central value is $\log x_{\rm H_2} \approx -3.2$ afterwards. The main difference between the four runs is the redshift at which the gradients develop.

Figure 2.11 shows typical density, temperature, and relative H_2 and e^- abundance profiles across the sphere at redshift $z = 10$ for each of the four runs. The $x_{\rm H_2}$ profiles are almost flat at the outer part of the sphere, and well approximated by a power law $x_{\rm H_2} \propto r^{-0.4}$ within the post-shock region, while the e^- fraction profiles behave less regularly. Note that the power-law profile for the density $\rho \propto r^{-2.25}$ predicted by the self-similar solution (Bertschinger 1985) is obtained near the center (see § 2 and the thick line in panel 11a).

Figure 2.12 shows the fraction of the collapsed baryonic mass out of the total bound baryonic mass vs. redshift for the runs 5-8, for two different values of the

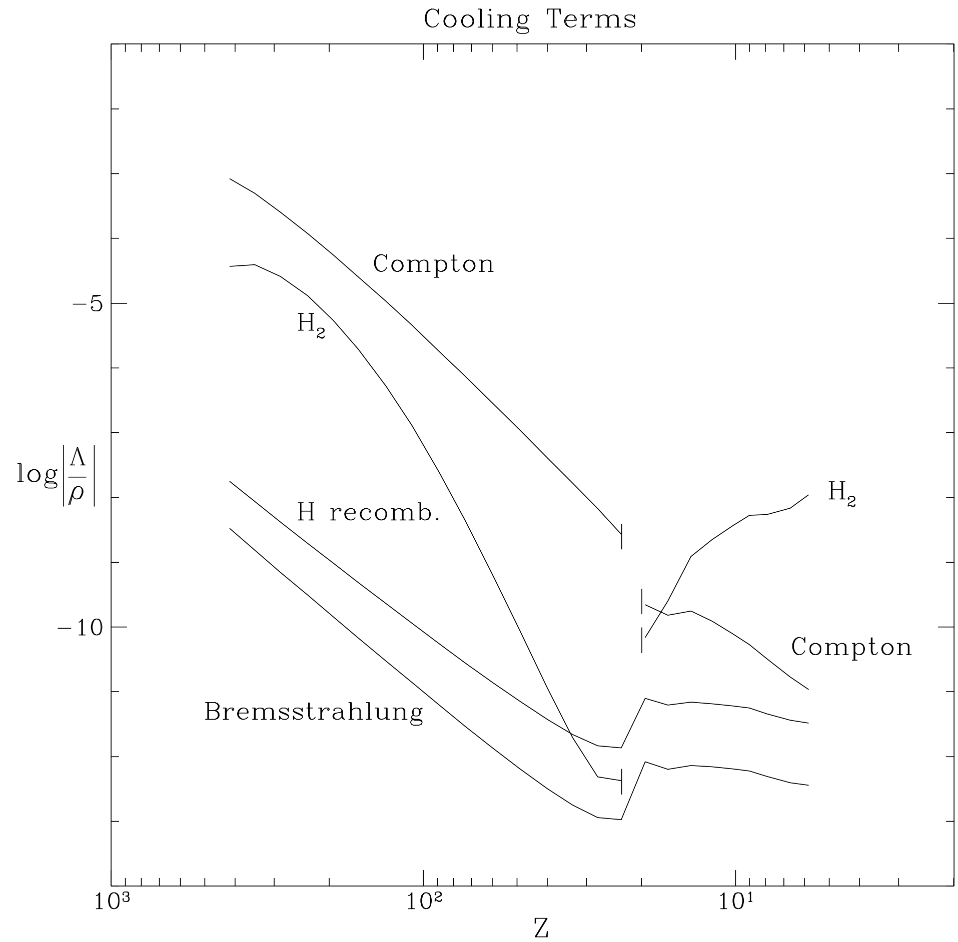


Figure 2.9: *The redshift evolution of the four most efficient radiative cooling/heating terms within the shell enclosing* 12% *of the total baryonic mass* $M_{\rm bound}$ *in a typical run (number 6). Before the gas temperature rises above the CBR temperature, both Compton scattering and the radiative transitions of* $\rm H_2$ *heat the gas.* Λ *changes sign for these processes in-between the tick marks. After* $T_{\rm b} > T_\gamma$ *(or* $z \lesssim 22$*), both of these mechanisms cool the gas; the most efficient post-shock cooling is due to* $\rm H_2$*. The units of* Λ/ρ *are erg* $\rm s^{-1}g^{-1}$.

overdensity threshold for forming objects $\bar{\delta}_{\rm thr}$. The collapsed baryonic mass versus redshift curves are rather insensitive to the value of $\bar{\delta}_{\rm thr}$, especially for the objects that collapse early.

2.4.5 Collapse onto a black hole

Next, we perform the same runs as described in §4.4, but with a different central boundary condition; we replace the reflecting hard sphere with a black hole sink of a fixed radius. We choose the radius of the sink to be a small fraction of the initial radius of the first shell. The shells which cross the black hole radius are assigned arbitrary values of density and temperature, and are subsequently ignored. Under these conditions, the dark matter component does not bounce back and virialize since it is absorbed by the central black hole. Due to the post-shock shell crossing, the

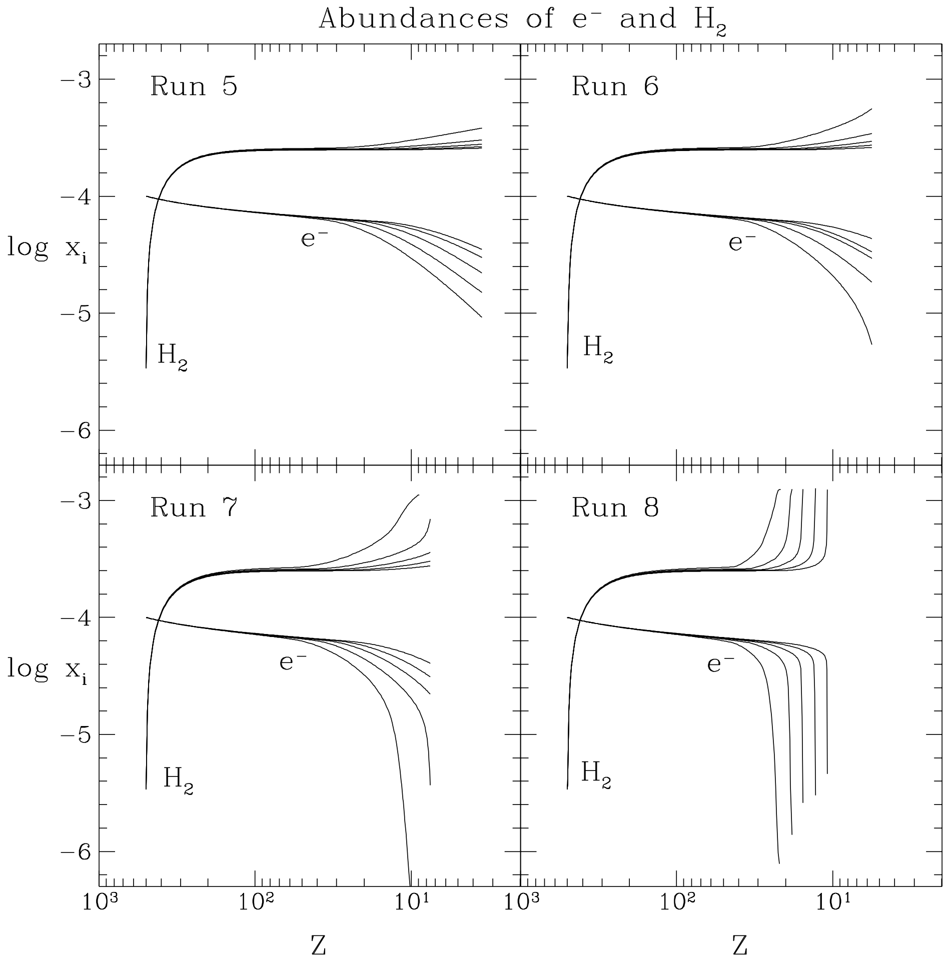


Figure 2.10: *Evolution of the relative abundances of H_2 and e^- for 5 shells (enclosing 2, 7, 12, 17, and 23% of the total baryonic mass $M_{\rm bound}$) in the runs numbered 5-8.*

gas component continues to collapse after it shocks, unlike in the case of a reflecting sphere. This leads to a continuous rise in the gas temperature after the shock, enabling more efficient post-shock cooling.

In the presence of a central black hole, we have found an oscillatory instability of the shock front in the cases involving efficient radiative cooling. Instead of propagating outwards monotonously as in the cases shown in Figure 2.7, the shock front in these cases can reverse its motion and oscillate. However, for the present application, the details of these oscillations can be glossed over, as we are only interested in the question whether shells reach $\bar{\delta} = \bar{\delta}_{\rm thr}$. The oscillation of the shock front occurs only for shells that have already collapsed according to our definition, hence this instability affects only the structural details of the objects we find but not the question of whether they form.

2.4.6 Discussion

We next explore the ($M_{\rm cloud}$, $\delta_{\rm dm,i}$) parameter space for the two different boundary conditions. In each case, we determine whether radiative cooling affects the dynamics of the gas and whether an object forms.

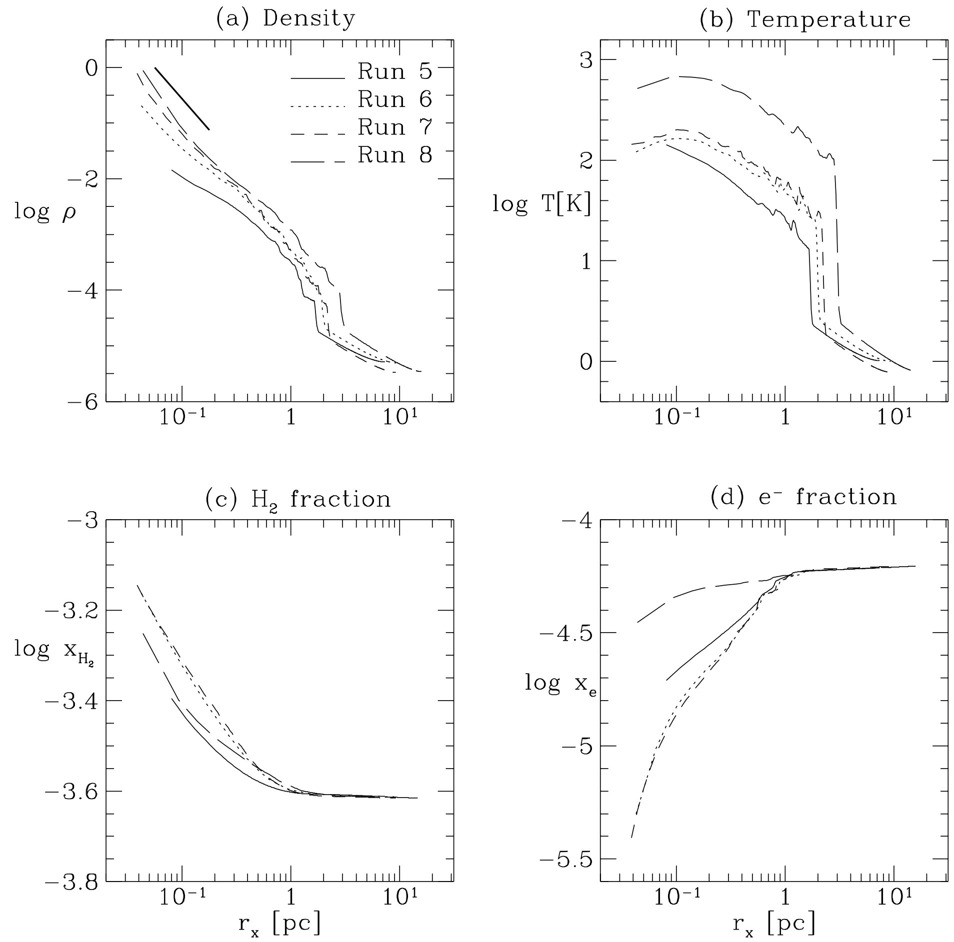


Figure 2.11: *Density, temperature, relative* H_2 *abundance and* e^- *abundance profiles in runs 5-8 at redshift* $z = 10$*. The abscissa is the comoving radius* $r_x = r(1+500)/(1+z)$*, where* r *is the physical radius. The density* ρ *is given in units of* $M_\odot/\mathrm{pc}^3$*. The thick line on the top left corner of panel (a) shows the slope of the density profile in the pressureless self-similar solution (Bertschinger 1985).*

To assess the significance of cooling in each run, we compared the trajectories of the gas shells with and without cooling. If the trajectories in the two cases were indistinguishable, we concluded that cooling had no effect; if they differed significantly, we concluded that cooling was efficient. To establish whether or not an object formed, we used a simple criterion: for an object to form, we required that at least half of the baryonic mass within the original cloud should collapse by a redshift $z = 10$, i.e., $M_{\rm obj}(z = 10)/M_{\rm cloud} \geq \frac{1}{2}$. This corresponds to 1/6 of the total bound mass ($M_{\rm bound}$) being above a density contrast of 200.

The main parameters which determine the post-shock cooling rate are the virial temperature $T_{\rm vir}$, and the abundance of molecular hydrogen $x_{\rm H_2}$ after the shock. While it is not possible to calculate $x_{\rm H_2}$ analytically, the scaling of the virial temperature can be derived from the spherical top-hat model (White 1994). Using our runs with a central hard sphere to calibrate this relation for the baryons, we find that for

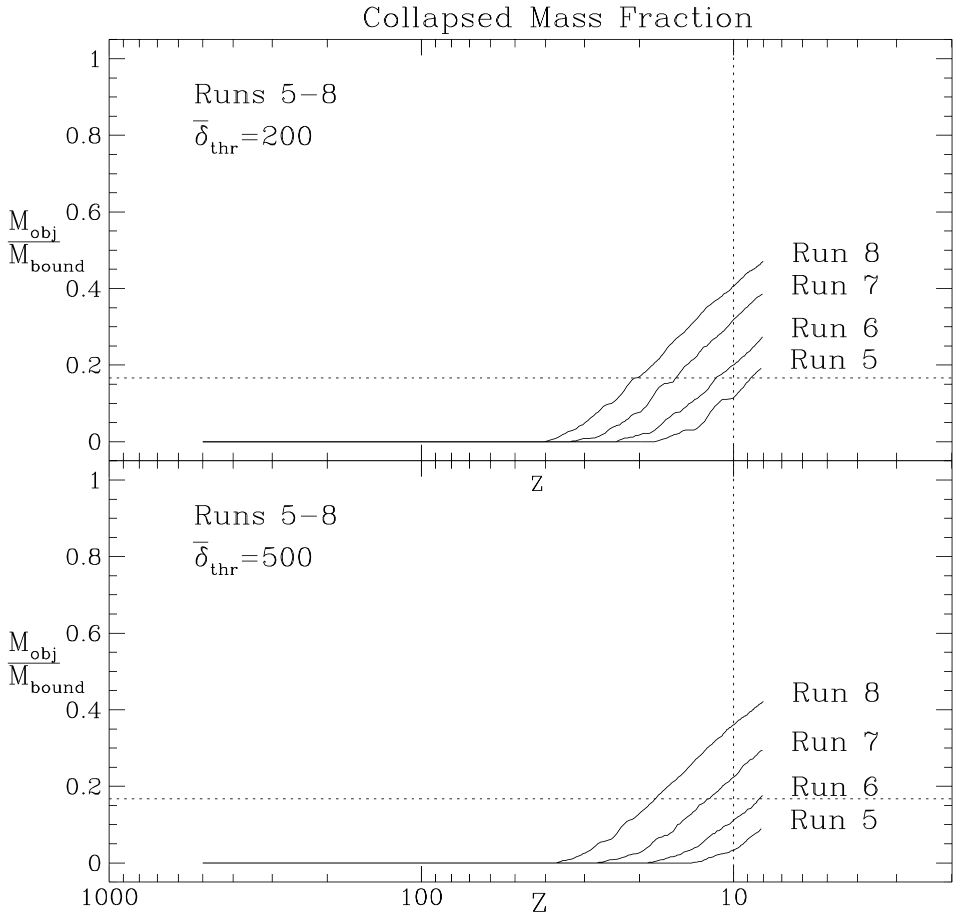


Figure 2.12: *Fraction of the total baryonic bound mass that has collapsed and formed an object in runs 5-8 vs. redshift. The upper panel shows results with* $\bar{\delta}_{\rm thr} = 200$*, the lower panel with* $\bar{\delta}_{\rm thr} = 500$*. The vertical and horizontal dashed lines illustrate the condition* $M_{\rm obj}/M_{\rm cloud} \geq 0.5$ *(or equivalently* $M_{\rm obj}/M_{\rm bound} \geq 0.17$*) at* $z = 10$*.*

$\Omega_0 = \Omega_{\rm b} + \Omega_{\rm dm} = 1$, the post-shock temperature of the gas shells closely follow

$$T_{\rm vir} = 55(1 + z_s)\left(\frac{h}{0.5}\right)^{2/3}\left(\frac{M_{\rm cloud}}{10^5 M_\odot}\right)^{2/3} \text{ K}, \tag{2.19}$$

where z_s is the shock redshift of each gas shell, and $M_{\rm cloud}$ is the baryonic mass of the cloud. The post-shock values for $x_{\rm H_2}$ do not vary substantially between runs, as shown in Figure 2.10.

In Figure 2.13 we show the grid of points we explored in the $(M_{\rm cloud}, \delta_{\rm dm,i})$ parameter space. The solid lines correspond to lines of constant central virial temperatures, $T_{\rm vir} = 85$ K and $T_{\rm vir} = 120$ K. The solid circles on this grid represent cases where cooling had no effect; the open circles represent cases where cooling was very efficient; and the open triangles represent intermediate cases, where cooling had a small but noticeable effect. The $(M_{\rm cloud}, \delta_{\rm dm,i})$ plane can be divided into two regions: the region where cooling is important, and the region where cooling is not important. According to our results, this division follows roughly a line of constant virial temperature $T_{\rm vir} \sim 10^2$ K.

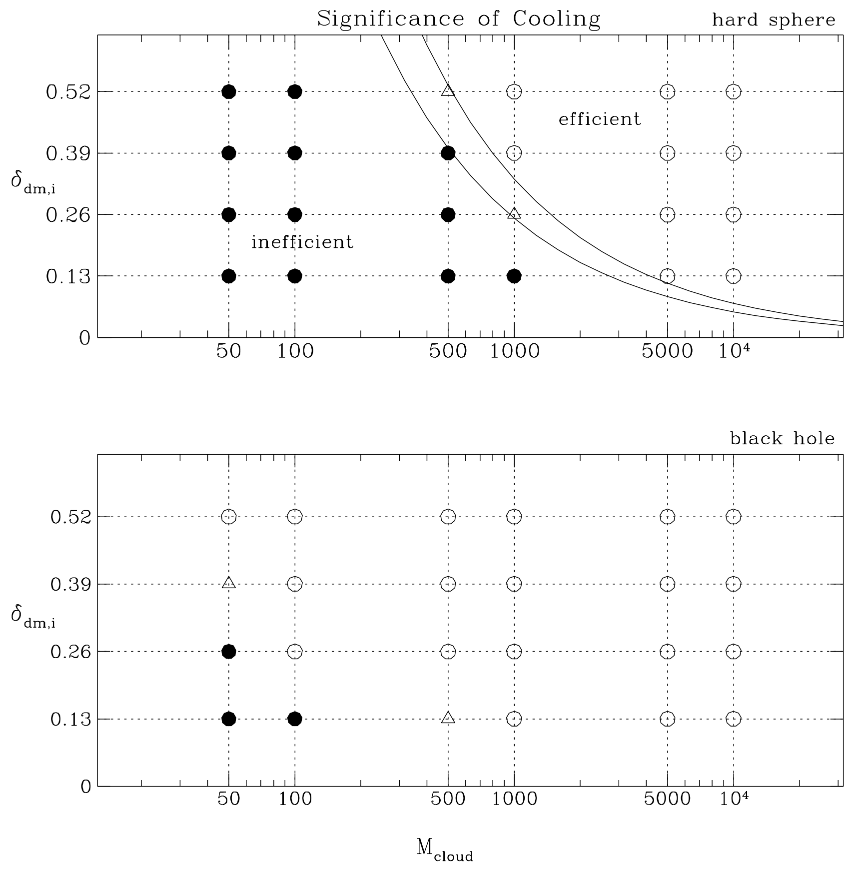


Figure 2.13: *Effectiveness of radiative cooling as a function of the two parameters* $\mathrm{M_{cloud}}$ *and* $\delta_{\rm dm,i}$. *Filled circles represent cases where cooling had no effect; open circles denote cases where cooling was very efficient; open triangles represent intermediate cases where cooling had minor but noticeable effect. On these mass-scales,* $\delta_{\rm dm,i} = 0.13$ *corresponds to (1.1-1.4)*σ *for the COBE-normalized (*$\sigma_8 = 1.22$*) cold dark matter power-spectrum (cf. Table 2.1). The upper panel refers to runs with a reflecting sphere in the center, and the lower panel to runs with a central black hole. The solid curves are lines of constant virial temperature (cf. eq. 2.19); the upper and lower curves correspond to temperatures of* 120 K *and* 85 K, *respectively.*

In Figure 2.14 we show again the above grid of points, but here the solid dots represent cases where no object formed by $z = 10$; the open dots represent cases where objects formed by $z = 10$; and the triangles represent intermediate cases where objects are just starting to form at $z = 10$. Again, we can divide the $(M_{\rm cloud}, \delta_{\rm dm,i})$ plane into two regions, depending on whether objects form or not. The constant $T_{\rm vir}$ curves from Figure 2.13 are also overlayed on this plane. According to the upper panel, we see that the division based on cooling goes through points of somewhat higher $\delta_{\rm dm,i}$ than the division based on collapse in the case of a central reflecting sphere. In other words, the transition from "no collapse" to "collapse" takes place in a regime where cooling is ineffective. Thus, radiative cooling does not introduce new mass or overdensity scales that are able to collapse.

The above result can be understood physically as follows. In order for H_2 cooling to affect the dynamics, the gas temperature $T_{\rm b}$ must satisfy two conditions. First, in order for cooling (rather than heating) to occur, the gas temperature must be

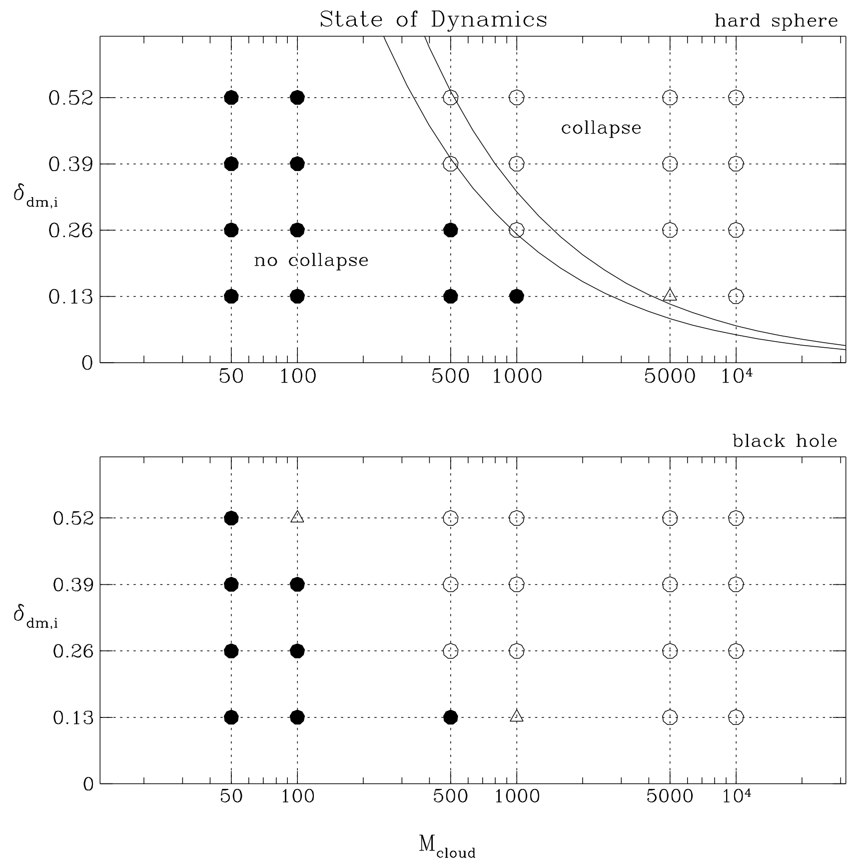


Figure 2.14: *Collapse as a function of the two parameters* $\mathrm{M}_{\rm cloud}$ *and* $\delta_{\rm dm,i}$. *The definition of "collapse" is given in § 4.2. Filled circles represent runs where none of the shells have collapsed by the redshift* $z = 10$*; open circles represent runs where more than half of the baryonic mass* $M_{\rm cloud}$ *has collapsed by* $z = 10$*; open triangles represent intermediate cases where only the innermost few shells have collapsed. The upper panel refers to runs with a reflecting sphere in the center, and the lower panel to runs with a central black hole. The solid lines are curves of constant virial temperature, as in Figure 2.13.'*

higher than the CBR temperature. Secondly, the H_2 cooling time needs to be smaller than the typical dynamical time; we find that this is only possible for a temperature $T_{\rm b} \gtrsim 10^2$ K. We find that the above two conditions cannot be satisfied simultaneously before the gas shocks. The gas temperature drops below 10^2K as early as $z \sim 100$, before any object forms, and rises above 10^2K again only after the gas shocks (cf. Fig. 2.8). However, by that late time the gas has already achieved an overdensity close to our collapse criterion $\bar{\delta} \sim 200$. Thus, even if cooling alters the post-shock trajectories significantly (as in run 8 in Fig. 2.7), its effect translates into a minor change in the collapsed baryonic mass as a function of redshift, $M_{\rm obj}(z)$.

The situation is somewhat different for collapses with a central sink (black hole) boundary condition. In the hard sphere case, the dark matter shells bounce back from the center, virialize, and stop moving subsequently. After a gas shell shocks, it is not crossed by dark matter-shells, unless radiative cooling is sufficiently effective to reduce its pressure support and allow it to fall towards the center. In the central sink case, the dark matter shells do not bounce back and virialize; instead, they are absorbed by the sink. Even after a gas shell shocks, it is crossed continuously by new dark matter shells. Therefore, gas shells continue to collapse after they shock even in the absence of radiative cooling. The continued infall heats the gas adiabatically, and the importance of post-shock cooling is enhanced relative to the hard-sphere case. The lower panel of Figure 2.13 shows that cooling is indeed effective for much smaller mass scales than in the hard-sphere case; the gas dynamics in clouds with a baryonic mass as low as $M_{\rm cloud} = 10^{2-3}{\rm M}_\odot$ is altered by cooling. For higher mass runs, H_2 cooling becomes so effective as to smooth-out the shocks; the gas shells fall-in rapidly and their trajectories show little evidence for a discontinuity. The cooling effects change the collapsed baryonic mass as a function of redshift, $M_{\rm obj}(z)$, and allow slightly lower mass-scales to collapse by $z = 10$. Indeed, the lower panel of Figure 2.14 shows that the transition from "no collapse" to "collapse" occurs for slightly lower masses than in the hard-sphere case, and in a regime where cooling is efficient.

2.5 Conclusions

In this paper we have investigated the early ($z \gtrsim 10$) formation of bound objects with masses in baryons comparable to the linear-theory Jeans-mass ($\sim 10^5{\rm M}_\odot$). We used a spherically symmetric Lagrangian code to analyze the dynamics of low-mass overdense regions in a cold dark matter cosmology. The code solved for the radial trajectories of the dark matter and gas shells, along with the density, temperature, and non-equilibrium abundances of various chemical ingredients of the gas in each shell.

The dynamics of a primordial gas cloud depends primarily on two parameters, the baryonic mass ($M_{\rm cloud}$) and the initial overdensity ($\delta_{\rm dm,i}$) of the cloud. We have computed the evolution of gas clouds on a grid of 48 runs, corresponding to different points in the ($M_{\rm cloud}$, $\delta_{\rm dm,i}$) parameter-space. We have done the calculations with two different types of central boundary conditions for the dark matter: a reflecting hard sphere, and a central sink (black hole).

Based on the above grid of runs, one can study the effect of radiative cooling by molecular hydrogen (H_2) on the dynamics of the baryons. We find that, in the hard sphere case, H_2 cooling is important only when the virial temperature of the collapsing object is $\gtrsim 100$ K (cf. eq. [2.19] and Fig. 2.13). However, this temperature regime is accessible only after shock-heating, when the gas has already condensed to be part of a nonlinear object. Before the gas enters the virialization shock, its temperature is lower than the CBR temperature and radiative H_2 cooling is not possible. This implies that contrary to previous suggestions, radiative cooling could not have triggered the formation of the first generation of bound objects in the universe.

However, the evolution of the gas after it shocks can be substantially altered due to H_2 cooling. This effect is particularly emphasized in the case of a central-sink boundary condition (cf. lower panel of Fig. 2.13), where the dark matter shells cross the baryonic shells and pull them effectively towards the center. This process brings the baryonic shells to a thermodynamic state that favors efficient cooling.

In general, we find that the presence of shell crossing by dark matter allows baryonic objects with masses well below the linear-theory Jeans mass to form. Figures 2.7 and 2.14 demonstrate that the cores of objects containing $\lesssim 10^3 M_\odot$ in baryons can collapse in a standard CDM cosmology by a redshift $z = 10$. It is therefore possible that high-mass stars had formed in the intergalactic space directly as a result of the primordial cosmological perturbations, and not just due to the commonly discussed process of fragmentation inside more massive objects. Intergalactic stars may have left their signature on the metalicity of the intergalactic medium and of Lyα absorption clouds through supernovae at high redshifts. It is difficult to detect the light from intergalactic stars unless they are assembled into high surface-brightness objects. However, such stars should result in gravitational microlensing events of quasars, with a lensing probability comparable to their cosmological density parameter $\sim \Omega_\star$ (Press & Gunn 1973).

The most significant limitation of our approach is the assumption of spherical symmetry. This assumption is forced upon us by practical computational considerations. Spherical symmetry may, in fact, apply to the collapse of a small subset of systems in environments with unusually low shear (Eisenstein & Loeb 1995b). However, most three-dimensional collapses are likely to be triaxial (Eisenstein & Loeb 1995a, and references therein). The early pancaking along the short axis would enhance the gas density and the corresponding H_2 formation, and could lead to a more efficient cooling of the gas. Fragmentation may lead to similar effects (see Ruzmaikina 1973, and Kashlinsky & Rees 1983). Future three-dimensional simulations are essential in exploring this more complicated dynamics.

Chapter 3

H_2-Cooling of Primordial Gas Triggered by UV Irradiation

Zoltán Haiman[1,2], Martin J. Rees[2], and Abraham Loeb[1]

[published in *The Astrophysical Journal*, **467**, 522 (1996)]

ABSTRACT

We investigate the formation of molecular hydrogen (H_2) in a primordial H+He gas cloud irradiated by a power-law UV flux. We find that at high densities ($\gtrsim 1$ cm^{-3}) and low temperatures ($\lesssim 10^4$ K), the background radiation enhances the formation of H_2 and results in molecular cooling dominating over photoionization heating. This process could accelerate the collapse and fragmentation of dense objects at high redshifts.

3.1 Introduction

It has long been hypothesized, based on studies of QSO absorption spectra (Gunn & Peterson 1965), that the intergalactic gas is photoionized by an intense UV background. Indeed, a background flux level $\sim 10^{-21\pm1}$ erg cm^{-2} Hz^{-1} sr^{-1} at the Lyman limit has been inferred from the proximity effect in the vicinity of QSOs at redshifts $z \approx 2-4$ (see Bechtold 1994, and references therein). This UV background could have resulted from either QSO or stellar emission (e.g., Miralda-Escudé & Ostriker 1990; Meiksin & Madau 1993; Shapiro, Giroux, & Babul 1994; Giroux, Fardal, & Shull 1995). Aside from its impact on the ionization state of the universe, the UV background heats the intergalactic gas and therefore raises its Jeans mass. Couchman & Rees (1986) realized the important negative feedback that this process might

[1]Astronomy Department, Harvard University, 60 Garden Street, Cambridge, MA 02138, USA
[2]Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 1RZ, England

have on the formation of low-mass objects at high redshift. More recently, Efstathiou (1992) suggested that photoionization heating of the intergalactic medium to temperatures $\gtrsim 10^4$ K could have suppressed the formation of dwarf galaxies ($\sim 10^9 M_\odot$) at moderate redshifts, and Babul & Rees (1992) have argued that the formation epoch of faint blue galaxies was indeed delayed by this process until $z \lesssim 1$ (see also Babul & Ferguson 1996). The negative feedback that photoionization heating has on structure formation on small mass scales was verified recently by a number of numerical simulations (Hernquist et al. 1996; Quinn, Katz, & Efstathiou 1996; Steinmetz 1995; Thoul & Weinberg 1995).

Naively, one would expect that a relatively cold gas cloud ($\lesssim 10^4$ K) could only heat due to its interaction with the ionizing background. However, this expectation ignores the formation of molecular hydrogen (H_2). In this paper we show that the increased abundance of free electrons due to photoionization tends to catalyze the formation of molecular hydrogen, which at sufficiently high densities results in a net *cooling* of the gas. To demonstrate this effect we consider a stationary homogeneous gas cloud which is being illuminated by an ionizing flux. §2 describes our method of calculation and §3 presents our results. Finally, we briefly sketch the potential cosmological implications from this effect in §4.

3.2 Method of Calculation

We consider a homogeneous gas cloud of primordial composition at a density $\rho = \mu m_{\rm p} n$ and temperature T, made up of the nine species H, H^-, H^+, He, He^+, He^{++}, H_2, H_2^+, and e^-. The chemical composition of the gas changes in time according to the rate equations

$$\frac{dn_{\rm i}}{dt} = \sum_{\rm l=1}^{9} \sum_{\rm m=1}^{9} a_{\rm lmi} k_{\rm lm} n_{\rm l} n_{\rm m} + \sum_{\rm j=1}^{9} b_{\rm ji} k_{\rm j} n_{\rm j}, \tag{3.1}$$

where $n_{\rm i}$ is the number density of the $i^{\rm th}$ species, $k_{\rm lm}$ denote reaction rate coefficients, $k_{\rm n}$ denote photo-ionization or photo-dissociation integrals, and the factors $a_{\rm lmi}, b_{\rm ni} = 0, \pm 1$, or ± 2, depending on the reaction. The full list of reactions and rate coefficients that we adopt are given in Appendix B.

We assume that the gas is illuminated by a uniform background radiation with a power-law flux extending up to 40keV,

$$J_\nu = J_{21} \left(\frac{\nu}{\nu_{\rm H}}\right)^{-\alpha} \times 10^{-21} \;\; {\rm erg\ s^{-1}\ cm^{-2}\ Hz^{-1}\ sr^{-1}} \qquad h\nu < 40{\rm keV}. \tag{3.2}$$

Unless stated otherwise, in our calculations we use the values $J_{21} = 0.1$ and $\alpha = 0.7$. This normalization of J_{21} is preferred by recent simulations of the Lyα forest (Hernquist et al. 1996; Miralda-Escudé et al. 1996) and is also compatible with the expected emission from the known QSO population (Meiksin & Madau 1993). The power-law index $\alpha = 0.7$ is in accord with quasar spectra, and is also consistent with the fit to the X-ray background spectrum up to around 40 keV (Fabian & Barcons 1992).

The incident spectrum in equation (3.2) characterizes the background flux before it enters the gas cloud. However, if the gas cloud is optically thick at the Lyman-limit, the shape of the spectrum will be modified due to H and He-photoionization as the radiation penetrates into the cloud. To mimic the effect of radiative transfer, we modify the spectrum to the form,

$$J_\nu = J_{21}\left(\frac{\nu}{\nu_{\rm H}}\right)^{-\alpha} \exp\{-N_{\rm H}\sigma_{\rm H}(\nu) - N_{\rm He}\sigma_{\rm He}(\nu)\} \times 10^{-21} \;\; {\rm erg\ s^{-1}\ cm^{-2}\ Hz^{-1}\ sr^{-1}}, \tag{3.3}$$

where $N_{\rm H}$ and $N_{\rm He}$ are the column densities (in units cm^{-2}) of neutral hydrogen and Helium through which the incident flux is processed. For simplicity, we assume that the incident radiation is effectively processed through a uniform slab of gas of length ℓ and define

$$N_{\rm i} = \bar{f}_{\rm i}\left(\frac{\rho}{\mu m_p}\right)\ell = 9.05\times 10^{20}{\rm cm}^{-2}\times \bar{f}_{\rm i}\left(\frac{T}{10^3{\rm K}}\right)^{\frac{1}{2}}\left(\frac{n}{1{\rm cm}^{-3}}\right)^{\frac{1}{2}}\mu^{-1}\times\frac{\ell}{\ell_{\rm Jeans}} \ . \tag{3.4}$$

Here $\bar{f}_{\rm i}$ are the average neutral hydrogen (i=H) and helium (i=He) fractions (by number) within the slab, μ is the mean molecular weight, n is the total proton number density, and $\ell_{\rm Jeans}$ is the Jeans length for the assumed density, temperature and composition of the slab.

The column density of molecular hydrogen H_2 in the slab determines the absorption by the Werner and Lyman bands of H_2 (see reaction (28) in Appendix B), which in turn shields the molecular hydrogen further inside the cloud against photodissociation. In a realistic cloud, the average H_2-fraction for a given $N_{\rm H}$ depends on the density and temperature profiles within the cloud. Since we do not model these profiles (except in Figure 3.2), $\bar{f}_{\rm H_2}$, as well as $\bar{f}_{\rm H}$ and $\bar{f}_{\rm He}$ in equation (3.4) must be specified a priori.

In addressing the question whether the gas cools or heats due to the radiation background at different densities or temperatures, we assume chemical equilibrium. This assumption makes our conclusions independent of the details of hydrodynamic and chemical history of the gas. (We examine below, as a separate question, the time-scale required for establishing the assumed chemical equilibrium). The chemical equilibrium abundances of the various species are solved by integrating equations (3.1) until the right-hand-side vanishes. Using these abundances, we then calculate the total cooling and heating functions, $\Lambda_{\rm cool}$ and $\Lambda_{\rm heat}$; these quantities are both positive and are expressed in units erg s^{-1} cm^{-3}. The interaction between the radiation background and the gas includes the following processes: photoionization and dissociation, Compton scattering, collisional ionization and excitation (including H_2), recombination, and Bremsstrahlung emission. Numerical expressions for these processes are listed in Appendix A.

3.3 Results and Discussion

Under chemical equilibrium and the fixed flux parameters ($J_{21} = 0.1, \alpha = 0.7$), the four parameters that define the state of the gas are its density ρ and temperature T, the neutral hydrogen column density of the shielding slab, $N_{\rm H}$, and its corresponding H_2 column density, $N_{\rm H_2}$. At each density, temperature, and H column density, the slab width ℓ can be expressed[3] in units of the Jeans length using equation (3.4).

[3]To do this conversion, $\bar{f}_{\rm H}$ must also be specified. However, as long as $N_{\rm H} \lesssim 10^{21}$ and $T \lesssim 10^4$, hydrogen is essentially neutral, so that $\bar{f}_{\rm H} = 0.927$ (the difference from unity is only because of

Table 3.1: *The equilibrium* H_2*-fraction* $f_{H_2,eq} = 2n_{H_2}/(n_H + n_{H-} + n_{H+} + 2n_{H_2} + 2n_{H_2+})$ *as a function of total number density* n*, temperature* T*, and column density* N_H*. For each column density* $(0 - 10^{21}$ cm$^{-2})$*, the five rows correspond to five different temperatures* $2.0 \le \log(T/\mathrm{K}) \le 4.0$*, the eight columns to the number densities* $-1.0 \le \log(n/\mathrm{cm}^{-3}) \le 6.0$*. The three tables show calculations for three different levels of self-shielding, (a): no self-shielding, i.e.* $N_{H_2}/N_H = 0$*, (b): intermediate self-shielding with* $N_{H_2}/N_H = 10^{-5}$*; (c): strong self-shielding with* $N_{H_2}/N_H = 0.05$.

(a)

0	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.36E-07	0.79E-06	0.10E-04	0.11E-03	0.83E-03	0.45E-02	0.18E-01	0.58E-01
2.5	0.66E-07	0.20E-05	0.29E-04	0.30E-03	0.22E-02	0.12E-01	0.46E-01	0.14E+00
3.0	0.10E-06	0.50E-05	0.82E-04	0.83E-03	0.61E-02	0.31E-01	0.11E+00	0.31E+00
3.5	0.99E-07	0.49E-05	0.41E-04	0.17E-03	0.56E-03	0.16E-02	0.49E-03	0.90E-04
4.0	0.28E-08	0.79E-07	0.64E-06	0.30E-05	0.11E-04	0.21E-04	0.26E-06	0.47E-07
10^{18}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.25E-06	0.30E-05	0.25E-04	0.14E-03	0.59E-03	0.21E-02	0.67E-02	0.21E-01
2.5	0.63E-06	0.85E-05	0.71E-04	0.39E-03	0.16E-02	0.55E-02	0.18E-01	0.56E-01
3.0	0.15E-05	0.24E-04	0.20E-03	0.11E-02	0.44E-02	0.15E-01	0.48E-01	0.14E+00
3.5	0.24E-05	0.29E-04	0.14E-03	0.50E-03	0.16E-02	0.27E-02	0.18E-03	0.30E-04
4.0	0.75E-07	0.62E-06	0.29E-05	0.11E-04	0.36E-04	0.16E-04	0.13E-06	0.35E-07
10^{19}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.40E-06	0.29E-05	0.13E-04	0.48E-04	0.16E-03	0.51E-03	0.16E-02	0.51E-02
2.5	0.11E-05	0.80E-05	0.36E-04	0.13E-03	0.42E-03	0.13E-02	0.43E-02	0.13E-01
3.0	0.32E-05	0.23E-04	0.99E-04	0.35E-03	0.12E-02	0.37E-02	0.12E-01	0.36E-01
3.5	0.82E-05	0.55E-04	0.22E-03	0.75E-03	0.21E-02	0.11E-02	0.41E-04	0.70E-05
4.0	0.94E-06	0.42E-05	0.16E-04	0.51E-04	0.11E-03	0.79E-05	0.92E-07	0.34E-07
10^{20}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.20E-06	0.11E-05	0.40E-05	0.13E-04	0.42E-04	0.13E-03	0.42E-03	0.13E-02
2.5	0.54E-06	0.29E-05	0.11E-04	0.35E-04	0.11E-03	0.36E-03	0.11E-02	0.35E-02
3.0	0.15E-05	0.81E-05	0.29E-04	0.96E-04	0.31E-03	0.97E-03	0.31E-02	0.96E-02
3.5	0.46E-05	0.24E-04	0.87E-04	0.28E-03	0.76E-03	0.31E-03	0.11E-04	0.18E-05
4.0	0.36E-05	0.15E-04	0.51E-04	0.12E-03	0.12E-03	0.64E-05	0.89E-07	0.33E-07
10^{21}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.69E-07	0.34E-06	0.12E-05	0.38E-05	0.12E-04	0.38E-04	0.12E-03	0.38E-03
2.5	0.19E-06	0.92E-06	0.31E-05	0.10E-04	0.32E-04	0.10E-03	0.32E-03	0.10E-02
3.0	0.52E-06	0.25E-05	0.86E-05	0.28E-04	0.88E-04	0.28E-03	0.88E-03	0.28E-02
3.5	0.17E-05	0.78E-05	0.26E-04	0.83E-04	0.22E-03	0.90E-04	0.31E-05	0.52E-06
4.0	0.45E-05	0.20E-04	0.64E-04	0.13E-03	0.11E-03	0.63E-05	0.89E-07	0.33E-07

Table 3.1(b)

0	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.36E-07	0.79E-06	0.10E-04	0.11E-03	0.83E-03	0.45E-02	0.18E-01	0.58E-01
2.5	0.66E-07	0.20E-05	0.29E-04	0.30E-03	0.22E-02	0.12E-01	0.46E-01	0.14E+00
3.0	0.10E-06	0.50E-05	0.82E-04	0.83E-03	0.61E-02	0.31E-01	0.11E+00	0.31E+00
3.5	0.99E-07	0.49E-05	0.41E-04	0.17E-03	0.56E-03	0.16E-02	0.49E-03	0.90E-04
4.0	0.28E-08	0.79E-07	0.64E-06	0.30E-05	0.11E-04	0.21E-04	0.26E-06	0.47E-07
10^{18}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.25E-06	0.31E-05	0.26E-04	0.15E-03	0.61E-03	0.21E-02	0.70E-02	0.22E-01
2.5	0.64E-06	0.86E-05	0.72E-04	0.40E-03	0.16E-02	0.56E-02	0.18E-01	0.57E-01
3.0	0.15E-05	0.24E-04	0.20E-03	0.11E-02	0.44E-02	0.15E-01	0.49E-01	0.14E+00
3.5	0.24E-05	0.29E-04	0.14E-03	0.50E-03	0.16E-02	0.27E-02	0.18E-03	0.30E-04
4.0	0.75E-07	0.62E-06	0.29E-05	0.11E-04	0.36E-04	0.16E-04	0.13E-06	0.35E-07
10^{19}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.55E-06	0.39E-05	0.18E-04	0.66E-04	0.22E-03	0.70E-03	0.22E-02	0.70E-02
2.5	0.14E-05	0.97E-05	0.43E-04	0.15E-03	0.51E-03	0.16E-02	0.51E-02	0.16E-01
3.0	0.36E-05	0.25E-04	0.11E-03	0.39E-03	0.13E-02	0.41E-02	0.13E-01	0.40E-01
3.5	0.86E-05	0.57E-04	0.23E-03	0.78E-03	0.22E-02	0.11E-02	0.41E-04	0.70E-05
4.0	0.94E-06	0.42E-05	0.16E-04	0.51E-04	0.11E-03	0.79E-05	0.92E-07	0.34E-07
10^{20}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.20E-05	0.11E-04	0.41E-04	0.14E-03	0.44E-03	0.14E-02	0.44E-02	0.14E-01
2.5	0.26E-05	0.14E-04	0.51E-04	0.17E-03	0.54E-03	0.17E-02	0.54E-02	0.17E-01
3.0	0.40E-05	0.21E-04	0.77E-04	0.25E-03	0.80E-03	0.25E-02	0.80E-02	0.25E-01
3.5	0.81E-05	0.42E-04	0.15E-03	0.48E-03	0.12E-02	0.33E-03	0.11E-04	0.18E-05
4.0	0.38E-05	0.16E-04	0.53E-04	0.13E-03	0.12E-03	0.64E-05	0.89E-07	0.33E-07
10^{21}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.47E-05	0.23E-04	0.80E-04	0.26E-03	0.81E-03	0.26E-02	0.81E-02	0.25E-01
2.5	0.92E-05	0.45E-04	0.15E-03	0.49E-03	0.16E-02	0.49E-02	0.15E-01	0.47E-01
3.0	0.17E-04	0.80E-04	0.27E-03	0.88E-03	0.28E-02	0.87E-02	0.27E-01	0.81E-01
3.5	0.30E-04	0.14E-03	0.45E-03	0.12E-02	0.11E-02	0.10E-03	0.31E-05	0.52E-06
4.0	0.13E-04	0.52E-04	0.13E-03	0.16E-03	0.11E-03	0.63E-05	0.89E-07	0.33E-07

We first focus on the sign of $\Lambda_{\rm tot} \equiv \Lambda_{\rm heat} - \Lambda_{\rm cool}$ as a function of these parameters. For reference, the equilibrium H_2-fractions as a function of density, temperature, and column density $N_{\rm H}$ are shown in Table 3.1. Figure 3.1 shows contours of $\Lambda_{\rm tot} = 0$ in the density-temperature plane for various values of neutral-H column densities ($N_{\rm H}$). At high temperatures ($T \gtrsim 10^4$ K), H_2 has no effect, and cooling always dominates through H-excitation. The prominent feature introduced into this figure by H_2 is the island of cooling below $T = 10^4$. Within this island, $\Lambda_{\rm heat}$ is dominated by H-photoionization heating, and $\Lambda_{\rm cool}$ is dominated by H_2-cooling; the net $\Lambda_{\rm tot}$ is negative.

In Figure 3.1(a) and Table 3.1(a), we assume $N_{\rm H_2} = 0$, i.e. we ignore self-shielding by H_2. Figures 3.1(b) and (c), and the corresponding Tables 3.1(b) and (c) show the effect of increasing self-shielding; in Figure 3.1(b) and Table 3.1(b) we

the presence of Helium). We make use of this conversion only in Figure 3.2, where this value of $f_{\rm H} = 0.927$ is indeed justified.

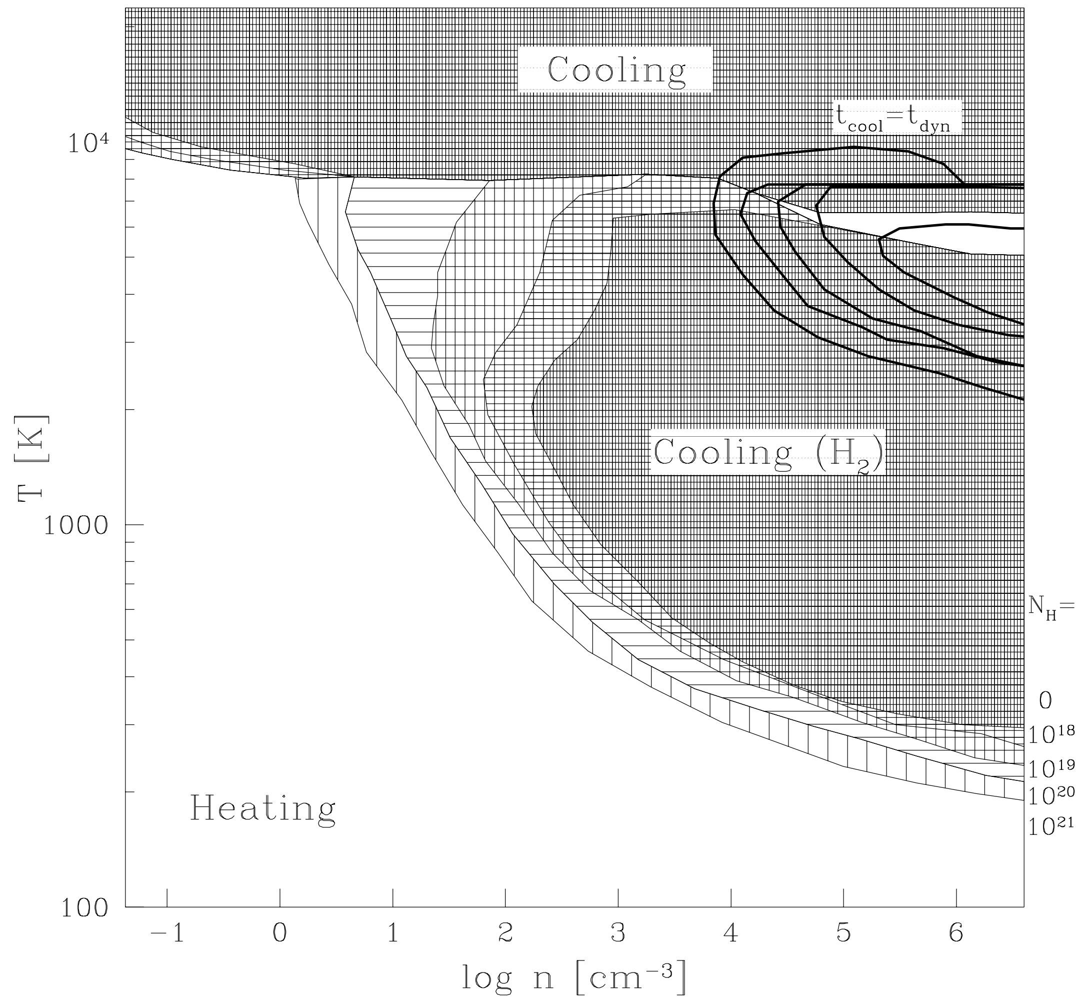
Table 3.1(c)

0	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.36E-07	0.79E-06	0.10E-04	0.11E-03	0.83E-03	0.45E-02	0.18E-01	0.58E-01
2.5	0.66E-07	0.20E-05	0.29E-04	0.30E-03	0.22E-02	0.12E-01	0.46E-01	0.14E+00
3.0	0.10E-06	0.50E-05	0.82E-04	0.83E-03	0.61E-02	0.31E-01	0.11E+00	0.31E+00
3.5	0.99E-07	0.49E-05	0.41E-04	0.17E-03	0.56E-03	0.16E-02	0.49E-03	0.90E-04
4.0	0.28E-08	0.79E-07	0.64E-06	0.30E-05	0.11E-04	0.21E-04	0.26E-06	0.47E-07
10^{18}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.58E-06	0.81E-05	0.95E-04	0.10E-02	0.10E-01	0.78E-01	0.34E+00	0.68E+00
2.5	0.15E-05	0.23E-04	0.26E-03	0.27E-02	0.25E-01	0.16E+00	0.50E+00	0.80E+00
3.0	0.35E-05	0.63E-04	0.72E-03	0.70E-02	0.59E-01	0.30E+00	0.66E+00	0.87E+00
3.5	0.41E-05	0.40E-04	0.17E-03	0.57E-03	0.18E-02	0.28E-02	0.18E-03	0.30E-04
4.0	0.76E-07	0.63E-06	0.29E-05	0.11E-04	0.37E-04	0.16E-04	0.13E-06	0.35E-07
10^{19}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.17E-04	0.21E-03	0.20E-02	0.15E-01	0.74E-01	0.25E+00	0.54E+00	0.80E+00
2.5	0.46E-04	0.55E-03	0.51E-02	0.36E-01	0.16E+00	0.44E+00	0.74E+00	0.90E+00
3.0	0.13E-03	0.15E-02	0.13E-01	0.85E-01	0.32E+00	0.65E+00	0.86E+00	0.95E+00
3.5	0.59E-04	0.25E-03	0.82E-03	0.25E-02	0.57E-02	0.13E-02	0.41E-04	0.70E-05
4.0	0.99E-06	0.43E-05	0.16E-04	0.51E-04	0.11E-03	0.79E-05	0.92E-07	0.34E-07
10^{20}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.19E-03	0.16E-02	0.81E-02	0.30E-01	0.93E-01	0.25E+00	0.52E+00	0.78E+00
2.5	0.50E-03	0.43E-02	0.21E-01	0.74E-01	0.21E+00	0.47E+00	0.74E+00	0.90E+00
3.0	0.13E-02	0.11E-01	0.53E-01	0.17E+00	0.42E+00	0.70E+00	0.88E+00	0.96E+00
3.5	0.23E-03	0.94E-03	0.30E-02	0.65E-02	0.41E-02	0.35E-03	0.11E-04	0.18E-05
4.0	0.47E-05	0.19E-04	0.61E-04	0.14E-03	0.12E-03	0.64E-05	0.89E-07	0.33E-07
10^{21}	-1.0	0.0	1.0	2.0	3.0	4.0	5.0	6.0
2.0	0.49E-03	0.26E-02	0.93E-02	0.29E-01	0.88E-01	0.23E+00	0.49E+00	0.75E+00
2.5	0.13E-02	0.70E-02	0.24E-01	0.74E-01	0.20E+00	0.45E+00	0.72E+00	0.89E+00
3.0	0.36E-02	0.19E-01	0.63E-01	0.18E+00	0.41E+00	0.69E+00	0.87E+00	0.96E+00
3.5	0.68E-03	0.29E-02	0.65E-02	0.46E-02	0.13E-02	0.10E-03	0.31E-05	0.52E-06
4.0	0.16E-04	0.63E-04	0.14E-03	0.16E-03	0.11E-03	0.63E-05	0.89E-07	0.33E-07

set $N_{\mathrm{H}_2}/N_{\mathrm{H}} = 10^{-5}$, and in Figure 3.1(c) and Table 3.1(c) we have set $N_{\mathrm{H}_2}/N_{\mathrm{H}} = 0.05$. The effect of self-shielding is to increase the equilibrium abundance of H_2, and to expand the boundaries of the cooling-island. In comparison with Figure 3.1(a), Figure 3.1(b) reveals that even a low molecular fraction (10^{-5}) causes a significant expansion of the boundary of the cooling island. The molecular fraction under realistic conditions is expected to be even higher (cf. Figure 3.2).

Note that on all three contour-plots, the island diminishes for a thin shielding slab, while its boundaries expand as the column density of the slab grows. The reasons for the inverse relation between the area of the H_2-cooling island and the total transmitted flux $F = \int J_\nu d\nu$ are twofold. The equilibrium H_2-fraction $f_{\mathrm{H}_2,\mathrm{eq}}$ is set by a balance between the two-step collisional formation process (either reaction [9] followed by reaction [10]; or reaction [7] followed by reaction [8]) and the photodissociation rate (either reaction [27] or reaction [28]). Hence, $f_{\mathrm{H}_2,\mathrm{eq}}$ increases with decreasing flux, and so does the H_2-cooling. In addition, photoionization heating decreases with decreasing flux. Roughly, the scalings are $\Lambda_{\mathrm{cool}} \sim 1/\sqrt{F}$ and $\Lambda_{\mathrm{heat}} \propto F$, so that

Figure 3.1(a)

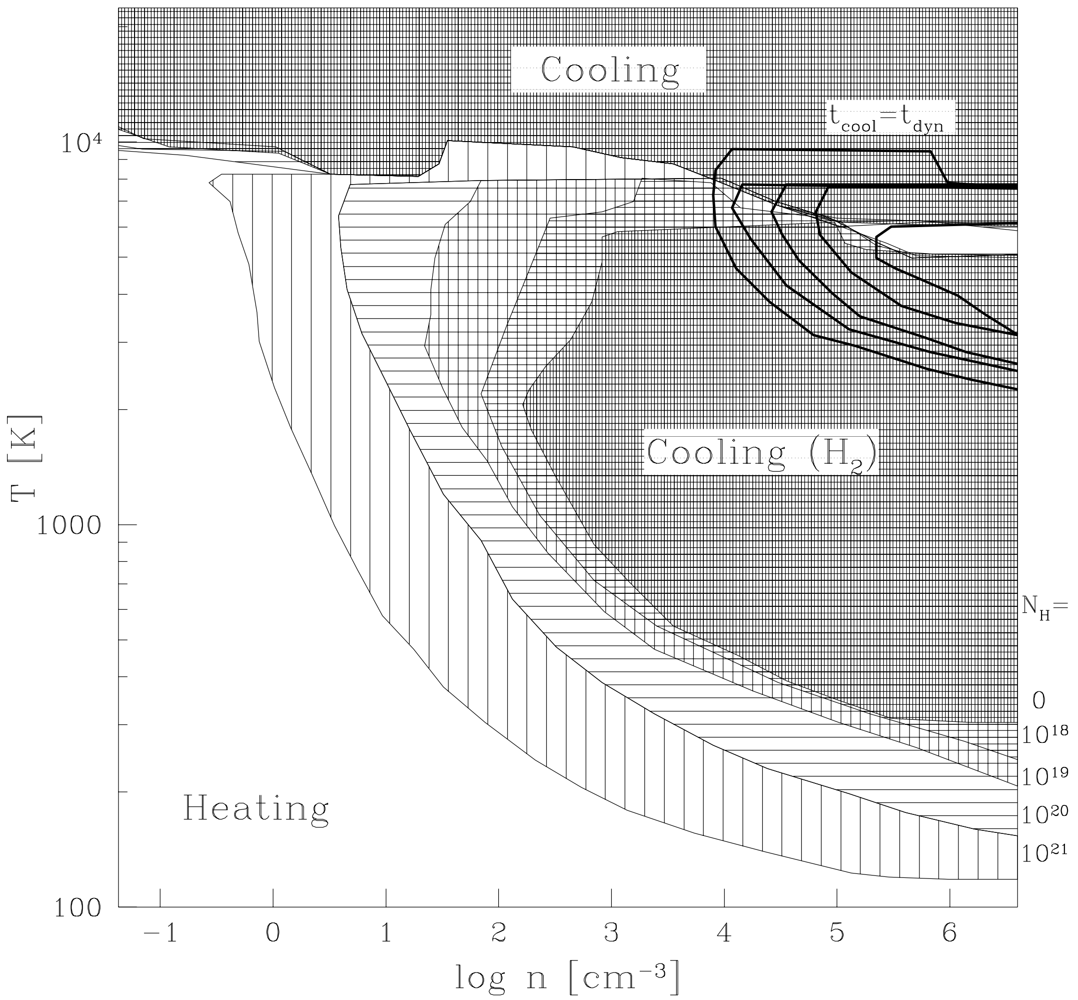


$\Lambda_{\rm cool}/\Lambda_{\rm heat} \sim F^{-3/2}$.

The exponential damping factor in equation 3.3 reduces the total flux, and causes the boundaries of the cooling-islands in Figure 3.1 to expand (note that a similar effect occurs if either the overall normalization, J_{21}, or the energy cutoff is lowered, or if the power-law index is increased). However, it should be borne in mind that as the ionizing flux is reduced, the formation rate of H_2 is lowered ($\propto \sqrt{F}$), and chemical equilibrium is less likely to be achieved within a Hubble time. In order to address this issue, we integrated equation (3.1) as a function of time and calculated the H_2 fraction after a Hubble time at $z = 5$, $f_{\rm H_2,Hub}$. Initially, the gas was assumed to be neutral and with no H_2, i.e. $n_{\rm H^-}=n_{\rm H^+}=n_{\rm He^+} = n_{\rm He^{++}}=n_{\rm H_2} = n_{\rm H_2^+} = n_{\rm e^-} = 0$ and $n_{\rm He}/n_{\rm H} = 0.078$.

Figure 3.2 shows profiles of $f_{\rm H_2,Hub}$, $f_{\rm H_2,eq}$ and the corresponding ratios of $t_{\rm cool}/t_{\rm dyn}$ across the shielding slab, for a fixed cloud density and temperature of $n = 1\ {\rm cm}^{-3}$ and $T = 10^3$ K. The profiles in this plot were constructed by subdividing the thick uniform slab into thin uniform plane-parallel slabs, and taking into account the

Figure 3.1(b)



continuous increase of optical depth in H, He, and H_2 from one thin slab to the next.

The upper panel demonstrates that as the optical depth of the shielding slab increases, the H_2-fraction initially grows slightly, while maintaining chemical equilibrium. The value of $f_{H_2,\mathrm{Hub}}$ peaks around ($\sim 10^{-3}$), at a slab thickness corresponding to $t_{\mathrm{chem}} = t_{\mathrm{Hub}}$, and declines at yet thicker slabs where it is out of equilibrium. The most important mechanism for the destruction for H_2 up to $N_{\mathrm{H}} \sim 10^{18-19}\mathrm{cm}^{-2}$ is direct photoionization (reaction 27). For larger N_{H}, the destruction rate is dominated by the two-step photo-dissociation (reaction 28), until the gas becomes self-shielding and collisional destruction by fast electrons dominates. The equilibrium state deep inside the cloud ($N_{\mathrm{H}} \sim 10^{23-24}$) is fully molecular, but this state is not reached within a Hubble time.

The lower panel of Figure 3.2 illustrates that the effectiveness of H_2-cooling is maintained at much greater optical depths than allowed by the requirement of chemical equilibrium. Indeed, the cooling time-scale remains much shorter than the dynamical time even as the slab thickness exceeds the Jeans length by two orders of magnitude. Even if H_2 is out of equilibrium, the increase in the optical depth can

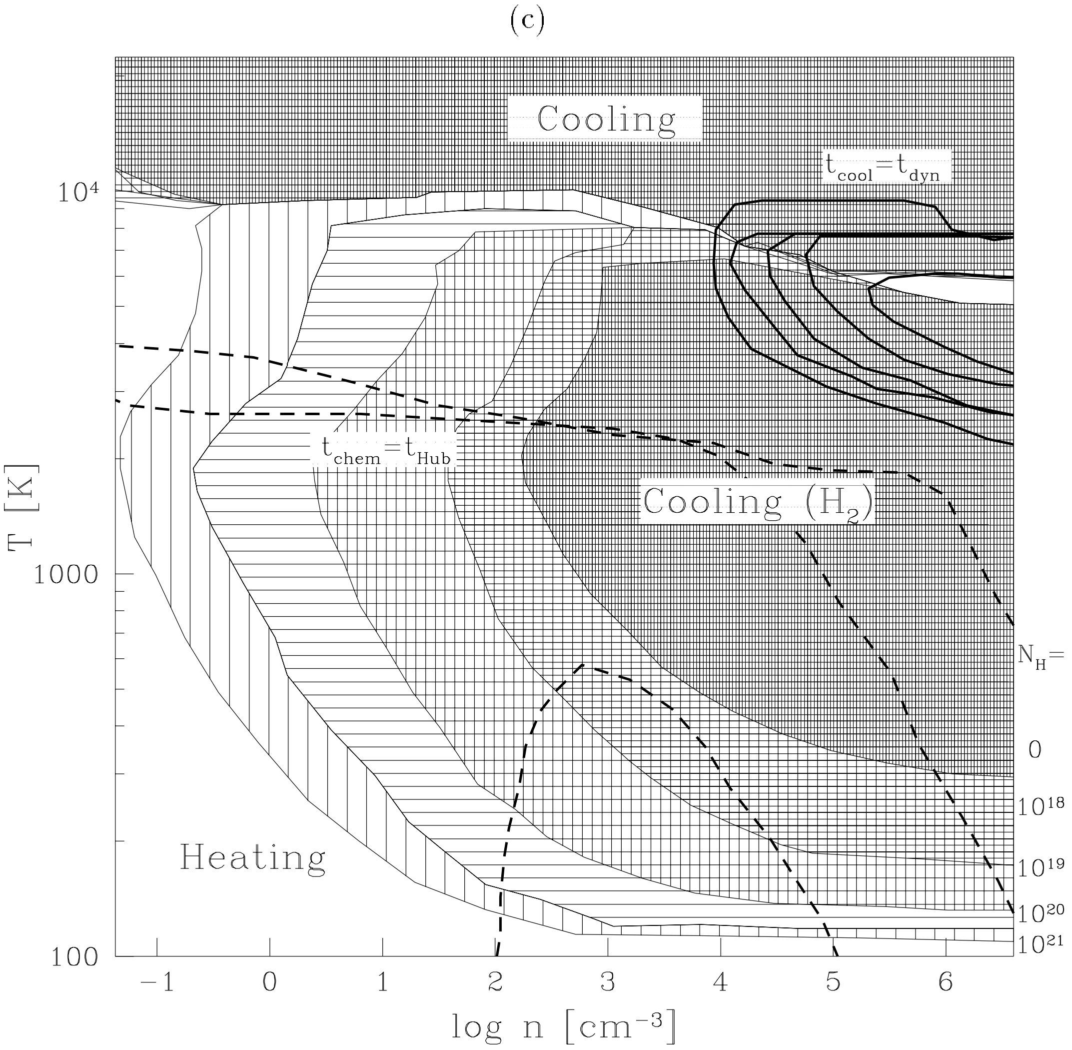


Figure 3.1: *Zero energy exchange contours ($\Lambda_{\rm heat} - \Lambda_{\rm cool} = 0$) for five different column-densities $N_{\rm H} = 0, 10^{18,19,20,21}$ cm^{-2}. The five solid lines in the upper right corner show corresponding lines of $t_{\rm cool} = t_{\rm dyn}$ (the right-most line corresponds to $N_{\rm H} = 0$, the left-most line to $N_{\rm H} = 10^{21}$) . Going to the left, the area enclosed by the lines defines increasingly larger regions of ineffective cooling ($t_{\rm cool} > t_{\rm dyn}$). The three panels show calculations for three different levels of self-shielding, (a): no self-shielding, i.e. $N_{\rm H_2}/N_{\rm H} = 0$, (b): intermediate self-shielding with $N_{\rm H_2}/N_{\rm H}$=10^{-5}; (c): strong self-shielding with $N_{\rm H_2}/N_{\rm H}$=0.05. The three dashed curves in (c) show lines of $t_{\rm chem} = t_{\rm Hub}$ for $N_{\rm H} = 10^{19}$ (lower curve) and $N_{\rm H} = 10^{20,21}$ (upper two curves). Above these lines, chemical equilibrium is established in a Hubble time. For $N_{\rm H} = 0$* and 10^{18}cm^{-2}*, the lines are off the plot.*

increase the ratio $\Lambda_{\rm cool}/\Lambda_{\rm heat}$ sufficiently to turn heating into cooling that is capable of affecting the gas dynamics. For comparison, Figure 3.2 shows how heating turns into cooling for three different normalizations, $J_{21} = 0.05, 0.10,$ and 0.15.

In conclusion, Figures 3.1 and 3.2 demonstrate that as the radiation field penetrates deeper into a gas cloud, it is more likely to cool the gas instead of heating it. There are, however, several limitations on the physical applicability of these results. In the rest of this section we discuss each of these limitations and demonstrate that they play only a minor role in most of the parameter space of the H_2-cooling island.

3.3.1 Cooling time vs. dynamical time

The contour-diagrams shown in Figure 3.1 provide only the sign of $\Lambda_{\rm heat} - \Lambda_{\rm cool}$. However, the fact that this quantity is negative within the H_2-cooling island does not in itself guarantee dynamical consequences. In order for cooling to effect the dynamics of a gas cloud, the cooling time needs to be shorter than the dynamical time. The curves of $t_{\rm cool} = t_{\rm dyn}$ are shown for five values of the column density, $N_{\rm H} = 0, 10^{18,19,20,21}$ on the density-temperature plane in Figure 3.1, where we adopt the following definitions for the cooling time,

$$t_{\rm cool} = \frac{(3/2)nk_{\rm B}T}{|\Lambda_{\rm heat} - \Lambda_{\rm cool}|} = \frac{(3/2)\rho k_{\rm B}T/\mu m_{\rm p}}{|\Lambda_{\rm heat} - \Lambda_{\rm cool}|} \tag{3.5}$$

and the dynamical time

$$t_{\rm dyn} = \frac{1}{\sqrt{6\pi G\rho}}. \tag{3.6}$$

It is apparent from these curves on the contour plots that most of the H_2-cooling island, except for its high-density high-temperature corner, lies in the region where $t_{\rm cool} < t_{\rm dyn}$. This implies that H_2-cooling in most cases is indeed efficient enough to affect the dynamics of the gas.

3.3.2 Chemical equilibrium vs. Hubble time

Figure 3.1 assumes chemical equilibrium. If the time-scale to reach chemical equilibrium, $t_{\rm chem}$, is too long, the H_2-fraction could be significantly lower than its equilibrium value $f_{H_2,\rm eq}$. Although the chemical reaction-network in equation (3.1) does not possess a single time-constant that characterizes relaxation to chemical equilibrium, the time-scale for H_2 to reach equilibrium is roughly proportional to the square of the number density of free electrons, i.e. $\propto 1/F$.

We compare $t_{\rm chem}$ and the Hubble time at $z = 5$, $t_{\rm H} = 2.0 \times 10^{16}$ s (assuming $h = 0.5$ and $\Omega = 1$). Although the chemical time-scale could also be compared to the dynamical time, this would be less relevant for virialized objects that do not undergo bulk motions in a dynamical time. Figure 3.1(c) shows (dashed) lines of $t_{\rm chem} = t_{\rm Hub}$ for three values of the column density, $N_{\rm H} = 10^{19,20,21}$. For $N_{\rm H} = 0$ and 10^{18}, H_2 is in equilibrium everywhere in Figure 3.1(c). Equilibrium is also reached everywhere in Figures 3.1(a) and (b). In the regions where $t_{\rm chem} < t_{\rm Hub}$, chemical equilibrium can, in principle, be established. In the regions of Figure 3.1(c) where $t_{\rm chem} > t_{\rm H}$, the H_2-fraction reached within a Hubble time $f_{H_2,\rm Hub}$ depends on the initial abundances and is always lower than the equilibrium abundance $f_{H_2,\rm eq}$ if initially f_{H_2}=0. Nevertheless,

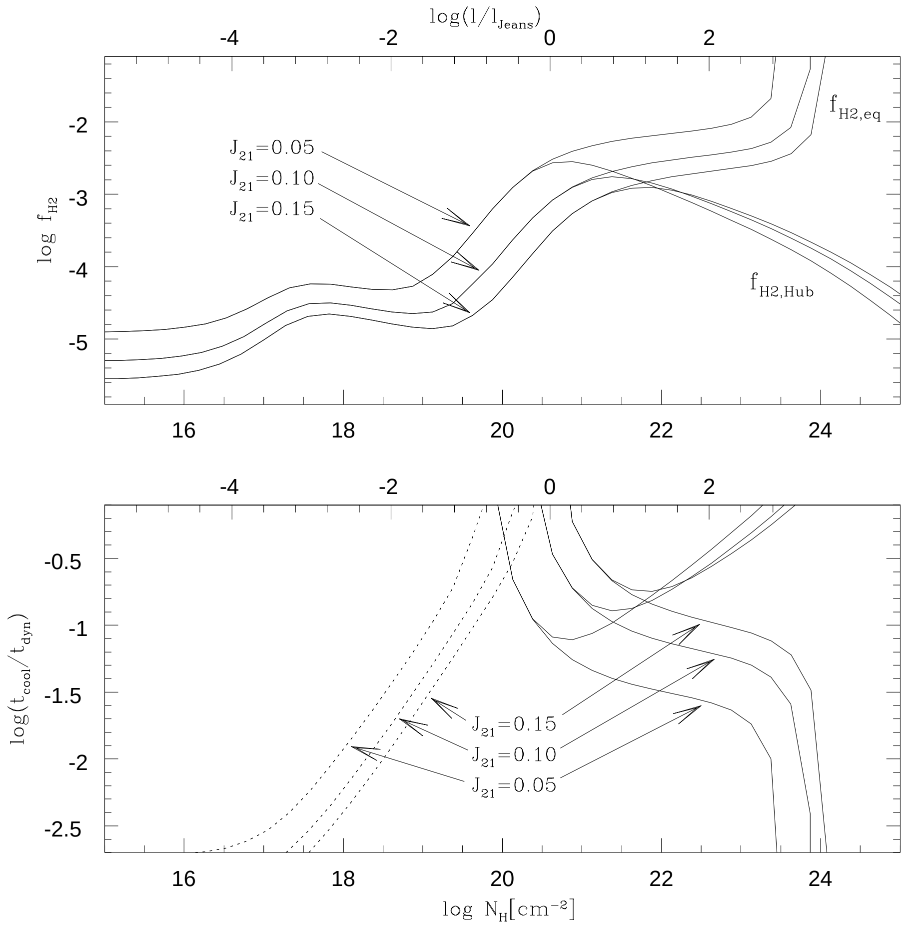


Figure 3.2: *Effects of shielding by a slab of column N_H (thickness ℓ) on the equilibrium* H_2*-fraction, and the* H_2*-fraction reached in a Hubble time (upper panel), and on the corresponding ratios of cooling time and the dynamical time (lower panel). In the lower panel, cooling is denoted by a solid line and heating by a dotted line. The gas is assumed to have a density* $1\ \mathrm{cm}^{-3}$ *and a temperature* $T = 10^3$ K*. The horizontal axis gives the column density* $\log N_{\mathrm{H}}$ *(bottom label) and the corresponding thickness of the shielding slab in units of the Jeans length (top label). Note that conversion between* $\log N_{\mathrm{H}}$ *and ℓ) is linear only up to* $\log N_{\mathrm{H}} \sim 10^{23}\mathrm{cm}^{-2}$*, i.e where hydrogen is neutral and correspondingly* $\bar{f}_{\mathrm{H}} \approx 0.927$ *is a constant.*

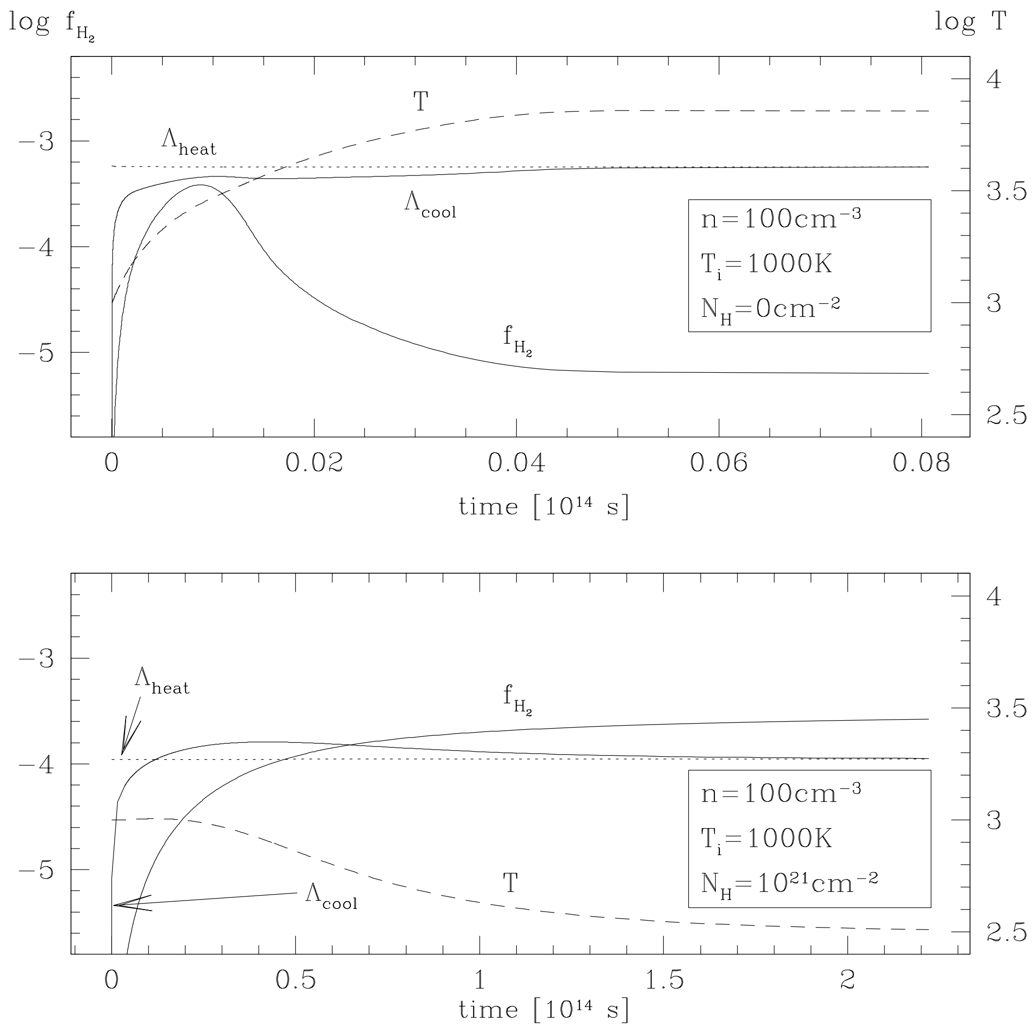


Figure 3.3: *The approach to equilibrium when thermal and chemical evolution are coupled. The vertical axis shows log of the* H_2*-fraction on the left, and log of the temperature in K on the right. The heating and cooling functions are arbitrarily normalized.*

we find that the contour-diagram Figure 3.1(c) is practically *unchanged* if $f_{H_2,\mathrm{Hub}}$ is used in place of $f_{H_2,\mathrm{eq}}$ to calculate Λ_{tot}. The out-of-equilibrium H_2 fraction becomes dynamically important, with $\Lambda_{\mathrm{tot}} < 0$ and $t_{\mathrm{cool}}/t_{\mathrm{dyn}} < 1$ within a Hubble time, even for the lowest column density shown in Figure 3.1, $N_{\mathrm{H}} = 10^{21}$ cm^{-2}.

3.3.3 Chemical equilibrium vs. thermal equilibrium

In realistic circumstances the approach to thermal and chemical equilibrium may be coupled. This coupling could affect the interpretation of Figures 3.1 and 3.2. Here, we will consider the regime where both the thermal and the chemical time-scales are shorter than the dynamical time. We therefore keep the density fixed, and consider the case $n = 100\mathrm{cm}^{-3}$.

If the initial H_2 abundance is small, then as soon as the background radiation turns on the gas will heat. The subsequent thermal evolution of the gas, however, depends on the competition between the photo-ionization heating rate and the increase in the H_2-cooling rate. Both the initial heating rate and the initial H_2-formation rate increase with increasing incident flux. However, the heating rate is proportional to the total flux F, while the H_2-formation rate grows only as $\sqrt{F}$. The H_2-formation rate is also sensitive to the gas temperature; it increases with temperature up to $\sim 3,000$K but then decreases abruptly at yet higher temperatures because of collisional dissociation. Thus, the relevant question is whether photoionization heating raises the gas temperature above $\sim 3,000$K before H_2 is formed in sufficient quantities to reverse the sign of $\Lambda_{\rm heat} - \Lambda_{\rm cool}$. As we show below, the answer to this question is positive for high incident fluxes and negative for low fluxes.

We solved the coupled time-evolution of the temperature and the relative abundances during the approach to thermal and chemical equilibrium. This was done by integrating equation (3.1), starting from $T = 10^3$ K and allowing the temperature and the rate coefficients $k_{\rm lm}$ to evolve at each timestep according to $\Lambda_{\rm heat} - \Lambda_{\rm cool}$. The initial abundances were chosen as in Figure 3.2. Figure 3.3 illustrates the sensitivity of the resulting temperature and H_2-fraction to the incident flux, characterized by the slab column density $N_{\rm H}$. For a higher flux ($N_{\rm H} = 0$, upper panel), $\Lambda_{\rm heat}$ remains roughly constant since the ionization fraction is low but $\Lambda_{\rm cool}$ is increasing as more H_2 is produced. When the temperature reaches $\sim 3,000$ K, molecular hydrogen is destroyed. The final equilibrium H_2-fraction is correspondingly low, $f_{\rm H_2,eq} = 6.0 \times 10^{-6}$, as the final equilibrium temperature reaches the high value of $7,160$ K. The lower panel shows a case when the incident flux is lower, due to a thick slab $N_{\rm H} = 10^{21}$. Here the H_2-fraction increases sufficiently rapidly so as never to allow the temperature to reach $\sim 3,000$ K. In fact, the temperature is raised only slightly before the cooling dominates the heating. In this case, the final equilibrium H_2-fraction is higher, $f_{\rm H_2,eq} = 3 \times 10^{-4}$, and the equilibrium temperature is low, 300 K. In fact, the equilibrium temperature in this case is *below* the initial temperature of $1,000$ K, implying that a net cooling feedback due to molecular hydrogen is possible.

3.4 Cosmological Implications

In summary, we find that a UV irradiation of a cold gas cloud with a density $\gtrsim 1$ cm^{-3} can trigger excess H_2-cooling on a time-scale shorter than its dynamical time. The effect is more pronounced in the inner cores of massive gas clouds which are shielded from photoionization heating by the surrounding material (cf. Fig. 3.1). Contrary to naive expectations, the existence of a strong ionizing background at redshifts $z \sim 2{-}4$, could have *accelerated* the collapse and fragmentation of dense objects, rather than increased their thermal pressure support. This H_2-cooling could have lowered the Jeans-mass threshold for fragmentation and led to enhanced star-formation rates in these objects. Since the UV background is found to decline to low values at low-redshifts (Kulkarni & Fall 1993), this enhancement in the fragmentation efficiency may have been a unique feature of the universe at redshifts $z \sim 2 - 4$.

The H_2-cooling effect may also have important implications on the collapse of objects, and the duration of the reionization epoch at higher redshifts. The cooling island in Figure 3.1(b) extends down to a density corresponding to a redshift $(1+z) \gtrsim 30\,(\Omega_{\rm b}h^2/0.025)^{-1/3}\,(1+\Delta/200)^{-1/3}$, where Δ is the overdensity in a collapsing region and $\Omega_{\rm b}$ is the baryonic density parameter. In particular if the very first objects to "light up" triggered the cooling effect discussed here, they could have accelerated the formation of additional ionizing sources, and reionization would have been completed abruptly. In contrast, if the first sources did not exert this positive feedback (e.g. because their UV spectrum was stellar and did not extend towards the X-ray band to trigger the formation of H_2), then the photoionization of the IGM must have taken longer with the Jeans mass rising continuously up to $10^{8-9}M_{\odot}$. These two scenarios differ in the amount of small-scale structure formed by the end of the reionization epoch.

Chapter 4

Destruction of Molecular Hydrogen During Cosmological Reionization

Zoltán Haiman[1], Martin J. Rees[2], and Abraham Loeb[1]

[published in *The Astrophysical Journal*, **476**, 458 (1997)]

ABSTRACT

We investigate the ability of primordial gas clouds to form and retain molecular hydrogen (H_2) during the initial phase of the reionization epoch. We find that before the Strömgren spheres of the individual ionizing sources overlap, the UV background below the ionization threshold is able to penetrate large clouds, and suppress their H_2abundance. The consequent lack of post-shock H_2cooling prevents the collapse and fragmentation of clouds with masses less than $\sim 10^9 M_\odot$. This negative feedback on structure-formation arises from the very first ionizing sources, and precedes the feedback due to the photoionization heating.

4.1 Introduction

Lately, there has been a renewed interest in the reionization of the intergalactic medium (IGM) due to its signature on the cosmic microwave background anisotropies (Tegmark, Silk & Blanchard 1994; Aghanim et al. 1996; Loeb 1996). Observations of QSO spectra imply that the IGM has been highly ionized already by the redshift $z \approx 5$ (Schneider, Schmidt & Gunn 1991). Evidence for the presence of an

[1]Astronomy Department, Harvard University, 60 Garden Street, Cambridge, MA 02138, USA
[2]Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 1RZ, England

ionizing background radiation with the required magnitude is provided by the proximity effect (Bajtlik, Duncan & Ostriker 1988; Lu, Wolfe, & Turnshek 1991; Bechtold 1994 and references therein). The ionizing sources could have either been stars or quasars (Miralda-Escudé & Ostriker 1990; Cen & Ostriker 1993; Meiksin& Madau 1993; Shapiro, Giroux & Babul 1994; Fukugita & Kawasaki 1994).

In Cold Dark Matter (CDM) cosmologies the first generation of objects forms at redshifts 10-50. Fragmentation of these objects into stars would result in early reionization. Couchman & Rees (1986) have argued that the heating associated with the photoionization of the IGM would inevitably raise the gas temperature to $T \sim 10^4$K and increase the value of the cosmological Jeans mass well above its initial value of $\sim 10^5 M_\odot$. This effect would tend to suppress further baryonic structure formation until the nonlinear mass scale rises above the new Jeans mass. This negative feedback serves as a buffer that monitors the fraction of baryons converted into massive stars and the metalicity of the IGM as a result of early reionization (Fukugita & Kawasaki 1994), and it could in principle explain the nearly universal metalicity value of $\sim 1\%$ observed in Lyman-α absorption systems (Tytler & Fan 1994; Tytler et al. 1995; Cowie et al. 1995).

The ability of primordial gas clouds with virial temperatures $T_{\rm vir} \approx 10^{2-4}$ K to cool radiatively and fragment after they virialize depends strongly on the existence of molecular hydrogen in them (Haiman, Thoul & Loeb 1996, hereafter HTL96; Tegmark et al. 1997). If the existence of H_2 is required in order to produce ionizing sources such as stars or quasars, then it is important to consider the negative feedback that an ionizing background has on the H_2 abundance in the universe. As hydrogen molecules can be destroyed by photons below the Lyman limit (Stecher & Williams 1967) to which the universe is transparent before being ionized, it is possible that the feedback due to H_2 destruction would precede the feedback due to photoionization heating.

In the early phase of reionization, after the first few sources of ionizing radiation appear, the Strömgren spheres of the individual sources are well separated from each other. During this phase, the ionizing photons from each source are absorbed at the boundary of their local Strömgren spheres, leaving the bulk of the Universe neutral. However, the photons below the Lyman limit escape the Strömgren spheres and are able to dissociate H_2 molecules elsewhere. Ionizing sources can therefore affect the H_2 abundance at a large distance (Dekel & Rees 1987).

In a previous paper (Haiman, Rees & Loeb 1996; hereafter HRL96) we studied the effect of quasar-like ionizing sources (with a power law spectrum) on the H_2 cooling of gas clouds after the universe is reionized. We have found that in very dense clouds ($n \gtrsim 10^3$ cm^{-3}) the ionizing radiation enhances the formation of molecules, and results in a net cooling of the gas. Here we examine the effect of the background radiation on the H_2 molecules at lower densities, inside clouds that are collapsing when most of the universe is still neutral. In contrast with HRL96, we calculate the spectrum of the background radiation before the reionization epoch. The processed photon spectrum above the Lyman limit is strongly suppressed due to HI continuum absorption throughout the universe. Below the Lyman limit, the spectrum is modulated due to HI line absorption. By estimating the required sizes for clouds to self-shield and retain their H_2 molecules, we demonstrate that the H_2 molecules inside virialized clouds are easily destroyed before the universe gets reionized. The photodissociation is due to photons softer than the Lyman limit; our results are therefore insensitive to whether the first objects emit a thermal (O-B star) or non-thermal spectrum. The early destruction of H_2 may provide a negative feedback that suppresses collapse and fragmentation prior to the feedback due to photoionization heating of neutral hydrogen. These two feedback scenarios could differ appreciably in the redshift at which the IGM became ionized, the predicted number of population III stars at the end of the reionization epoch, and the scales of the aggregates in which they, and hence the first nucleosynthesis products, are formed.

The paper is organized as follows. In § 2 we calculate the characteristic spectrum of the ionizing radiation during the early phase of reionization, when the Strömgren spheres around individual sources do not overlap. In § 3 we use this spectrum to examine the question whether relatively weak fluxes of photons can penetrate large clouds and destroy H_2 in them. Finally, § 4 summarizes the cosmological implications of the feedback due to the destruction of H_2 molecules.

4.2 Spectrum of UV Background During the Early Phase of Reionization

We consider the first stage of the reionization epoch, when most of the universe is still neutral and the Strömgren spheres around individual sources do not overlap. In order to calculate the spectrum of the UV background outside the Strömgren spheres, we make four assumptions: (i) the ionizing sources turn on simultaneously at the redshift $z_{\rm on}$; (ii) the sources are uniformly distributed in space; (iii) the emitted spectral intensity is a power-law with an index α; and (iv) the source population maintains a constant comoving density. Assumption (iii) would be appropriate if the ionizing sources were mainly accreting black holes; our conclusions, however, depend primarily on the emission of photons in the 11.18-13.6 ev range, and therefore apply with little change if the radiation comes from O and B stars.

The spectrum of the background radiation is suppressed above the Lyman limit through absorption by the column of neutral hydrogen and helium between the source and the observer. Photons below the Lyman limit will freely escape the Strömgren spheres around their sources. Since H_2 is dissociated by line absorption into the Lyman and Werner lines of molecular hydrogen between 11.18-14.67 eV, special care is needed in calculating the spectrum in this range.

The cumulative flux (in units of erg cm^{-2} s^{-1} ster^{-1} Hz^{-1}) observed at the frequency ν and redshift $z_{\rm obs}$ due to sources with a comoving emission coefficient $j(\nu,z)$ (in units of erg cm^{-3} s^{-1} ster^{-1} Hz^{-1}) between the redshifts $z_{\rm obs}$ and $z_{\rm on}$ can be written as

$$J(\nu, z_{\rm obs}) = \frac{c}{H_0}\int_{z_{\rm obs}}^{z_{\rm on}} j(\nu_z, z)\frac{\exp\{-\tau(z)\}dz}{(1+z)^2(1+q_0 z)^{1/2}}, \tag{4.1}$$

where $\nu_z \equiv \nu(1+z)/(1+z_{\rm obs})$, and $\tau(z)$ is the optical depth due to neutral hydrogen and helium between the observer's redshift $z_{\rm obs}$ and the redshift z,

$$\tau(z) = \frac{c}{H_0}\int_{z_{\rm obs}}^{z}\frac{\kappa(\nu_z, z)dz}{(1+z)^2(1+q_0 z)^{1/2}}, \tag{4.2}$$

$$\kappa(\nu, z) = n_{\rm H}(z)\sigma_{\rm H}(\nu) + n_{\rm He}(z)\sigma_{\rm He}(\nu). \tag{4.3}$$

Above the hydrogen ionization threshold, equation (4.1) must be evaluated numerically. Below this threshold, the flux is given by the integral in equation (4.1), with $\tau = 0$ if we ignore absorption due to lines. Assuming $q_0 = 1/2$ and writing the emission coefficient as $j(\nu, z) = j_0(\nu/\nu_H)^{-\alpha}$, where ν_H is the hydrogen ionization energy, we obtain in this case

$$J(\nu, z_{\rm obs}) = \frac{cj_0}{H_0}\left(\frac{\nu}{\nu_H}\right)^{-\alpha}\frac{(1+z_{\rm obs})^{\alpha}}{1.5+\alpha}\left[\frac{1}{(1+z_{\rm obs})^{1.5+\alpha}} - \frac{1}{(1+z_{\rm on})^{1.5+\alpha}}\right]. \tag{4.4}$$

In reality, the spectrum below the ionization threshold is processed due to line absorption in the Lyman bands of neutral hydrogen. When the IGM is still neutral, the optical depth across the universe in the resonant Lyman lines is $\sim 10^6$, and so each photon gets absorbed at the redshift when its frequency equals one of the Lyman lines. Hence, photons observed to have a frequency ν at a redshift $z_{\rm obs}$ could have only been emitted from redshifts that are between the observer and an effective screen located some distance away. The redshift of the screen $z_{\rm screen}$ depends on $\nu_{\rm Lyi}$, the frequency of the nearest Lyman line above ν, through the resonance condition

$$\nu\frac{(1+z_{\rm screen})}{(1+z_{\rm obs})} = \nu_{\rm Lyi}. \tag{4.5}$$

We can incorporate the effect of line absorption by replacing the turn-on redshift $z_{\rm on}$ with the screen redshift $z_{\rm screen}$ in equation (4.4). This is equivalent to multiplying the flux in equation (4.4) by the modulating factor

$$f = \frac{1 - [(1+z_{\rm obs})/(1+z_{\rm screen})]^{1.5+\alpha}}{1 - [(1+z_{\rm obs})/(1+z_{\rm on})]^{1.5+\alpha}} \tag{4.6}$$

So far we have ignored the radiative decay of the excited hydrogen atoms. In reality, each absorbing screen will re-emit line photons. Lyα photons are just re-cycled until they redshift out of resonance, because their destruction rate by the 2s$\rightarrow$1s two-photon decay is much slower than the cosmological expansion [cf. equation 6.114 in Peebles (1993)]. Therefore, there is no net absorption at the redshift corresponding to the Lyα resonance. On the other hand, Lyβ and higher Lyman line photons are re-emitted as a sum of Lyα and other Balmer or lower line photons. Consequently, the screens corresponding to Lyβ and higher Lyman lines produce two effects: (i) they block the view of all sources beyond the screen redshift, and (ii) they cascade to the ground state by emitting Lyα and some other, lower-frequency, photons. This results in a sawtooth-modulation of the spectrum in the frequency range from Lyβ up to the Lyman limit, and in a discontinuous step in the spectrum below the Lyα frequency.

Although it is outside the range of the Lyman and Werner bands of molecular hydrogen, the flux just below the Lyα frequency effects the dissociation rate of H^- and H_2^+. To calculate this flux, we assume that each radiative excitation to the 3p state or higher is followed instantaneously by a decay to the ground state with the emission of exactly one Lyα photon (plus other lower energy photons which are not important for our analysis). In addition, we assume that the re-emitted Lyα photons immediately shift out of resonance due to cosmological expansion, and proceed to the observer without further scattering. Under these assumptions, the additional flux at the frequency ν below Lyα due to Lyα emission from all screens can be written as a sum of terms, each of which being similar to equation (4.4),

$$\Delta J(\nu, z_{\rm obs}) = \sum_{i=\beta,\gamma,\ldots}\frac{cj_0}{H_0}\left(\frac{\nu}{\nu_H}\right)^{-\alpha}\left(\frac{\nu_{\rm Lyi}}{\nu_{\rm Ly\alpha}}\right)^{1-\alpha}\frac{(1+z_{\rm obs})^{\alpha}}{1.5+\alpha}\left[\frac{1}{(1+z_{\alpha})^{1.5+\alpha}} - \frac{1}{(1+z_{\rm i})^{1.5+\alpha}}\right], \tag{4.7}$$

where z_α is the redshift at which the absorbing screens need to be located in order that their Lyα emission would appear at a frequency ν at the redshift $z_{\rm obs}$,

$$z_\alpha = \frac{\nu_{\rm Ly\alpha}}{\nu}(1 + z_{\rm obs}) - 1, \tag{4.8}$$

and $z_{\rm i}$ is the highest redshift at which the Lyi photons absorbed at z_α could have been emitted, i.e. the location of a Ly(i+1)-absorbing screen,

$$z_{\rm i} = \frac{\nu_{\rm Ly(i+1)}}{\nu_{\rm Lyi}}(1 + z_\alpha) - 1. \tag{4.9}$$

The factor $(\nu_{\rm Lyi}/\nu_{\rm Ly\alpha})^{1-\alpha}$ in equation (4.7) accounts for the conversion of a Lyi to a Lyα photon during absorption and re-emission.

Figure 4.1 shows, as an example, the processed spectrum at $z_{\rm obs} = 25$ assuming the sources turn on at $z_{\rm on} = 35$. The top panel shows that above the Lyman limit the flux is suppressed up to the keV energy range, and the bottom panel presents the sawtooth modulation below the Lyman limit, as well as the discontinuous jump at the Lyα frequency.

The assumption that the UV source distribution is homogeneous remains valid as long as the radius around the observer out to the nearest absorbing screen is larger than the clustering length of the sources. Clearly the assumption breaks down for the highest transitions in the Lyman series. However, our calculation is concerned with the molecular Lyman or Werner lines which are separated from the nearest Lyman line by redshift intervals that are typically $[\Delta z/(1+z)] \gtrsim 10^{-2}$, i.e. much larger than the expected clustering lengths at high redshifts.

One may wonder whether it is possible to detect the reionization spectrum directly. The sawtooth template presented in figure 4.1 has a characteristic structure that should survive until the present time, provided the breakthrough epoch during which the universe becomes transparent to lines is sudden. Following an abrupt breakthrough epoch (as expected for an unclustered source population), the spectrum below the Lyman limit would be preserved and would show up as a slight modulation on top of the contribution from more recent sources which dominate the observed infrared background. The frequencies at which this sawtooth modulation appears depends on the reionization redshift, and its detection would provide invaluable information about the reionization epoch. Unfortunately, the amplitude of the sawtooth modulation is small. We estimate this amplitude based on the DIRBE observations (see Bond 1995a for a review), and assuming a constant comoving density of ionizing sources that ionize the universe abruptly at $z = 5$ and produce $J_{21} = 0.1$ at $z = 2$. For these conditions the modulation amplitude is smaller than 10^{-6}, implying that direct detection of the reionization signal is not feasible.

4.3 Destruction of Molecular Hydrogen

We may now use the spectrum illustrated in Fig. 4.1 to derive the equilibrium H_2 fraction inside overdense regions in the universe. In particular, we would like to find out how massive should a gas cloud be in order to form sufficient H_2 to cool on its dynamical time-scale.

In calculating the H_2 fraction, we consider a homogeneous cloud of primordial H/He ratio with a density ρ and temperature T, which is composed of the nine

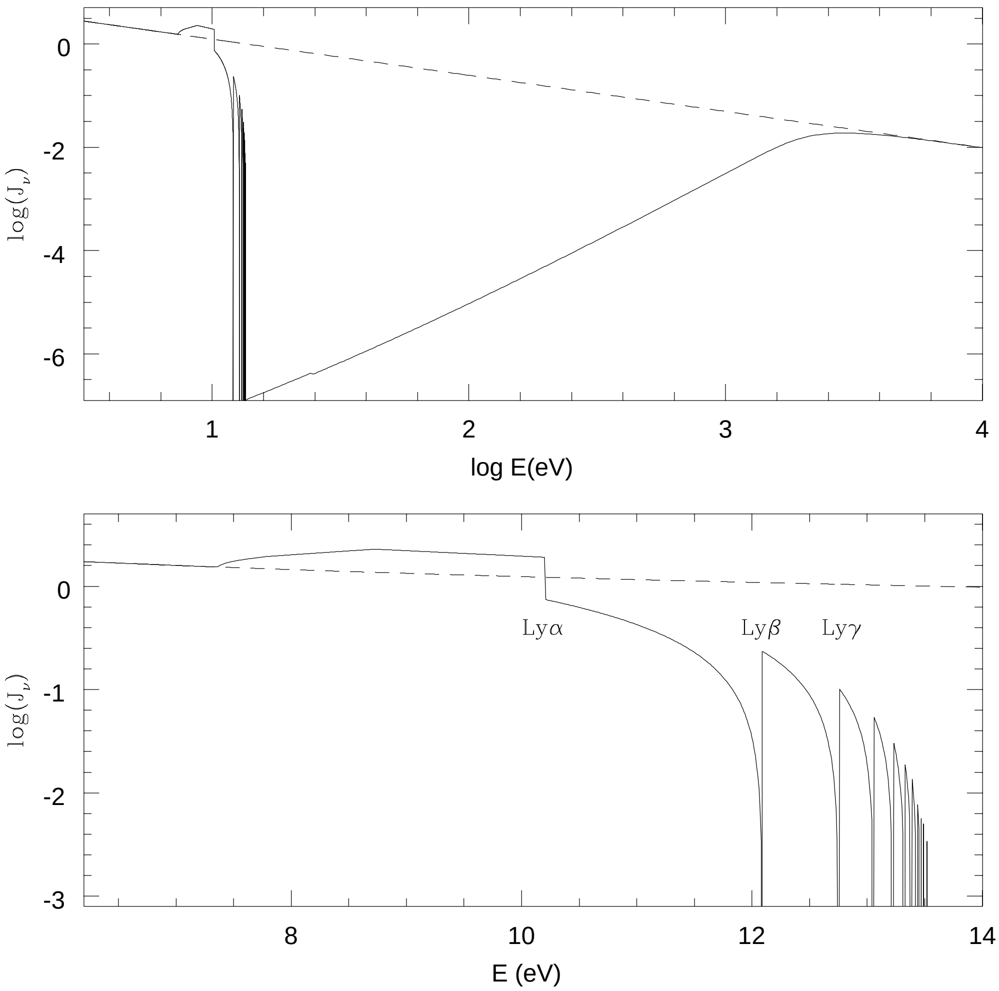


Figure 4.1: *The average flux (in units of* erg cm^{-2} s^{-1} ster^{-1} Hz^{-1}*) during the initial phase of the reionization epoch. The upper panel shows that absorption by neutral hydrogen and helium suppresses the flux above 13.6eV up to the keV range. The lower panel shows the sawtooth-modulation due to line absorption below 13.6eV.*

species H, H^-, H^+, He, He^+, He^{++}, H_2, H_2^+, and e^-. The chemical composition of the gas within an overdense region is determined by the interaction between these species and the background radiation. For the present purpose, we assume chemical and ionization equilibrium, which is reasonable for the densities and temperatures of interest. The equilibrium H_2 fraction is expected to provide an upper limit on the actual H_2 fraction and so our constraints on the destruction of H_2 are conservative. The assumption of equilibrium is powerful in that it makes our conclusions independent of the details of hydrodynamic and chemical history of the gas.

As emphasized before, the most important element of the chemical network in the present context is the dissociation of H_2 by photons with energies below 13.6eV. The significance of this two-step process was first highlighted by Solomon in 1965 (cf. Field et al. 1966) and subsequently studied quantitatively by Stecher & Williams (1967). Here we calculate the dissociation rate for this process by summing the oscillator strengths for the Lyman and Werner bands of H_2 from Allison and Dalgarno (1970), and multiplying it with the dissociation fractions (for the second step in the two-step process) from Dalgarno and Stephens (1970), and with the flux at each molecular line[3]. HTL96 have shown that the collapse of a primordial gas cloud beyond virialization depends on its ability to form H_2 and cool radiatively. The size of a cloud that is able to form H_2 depends, in turn, on its self-shielding in the Lyman and Werner lines. In order to correct the dissociation rate for self-shielding as a function of H_2 column density, we used the self-shielding function from de Jong, Dalgarno & Boland (1980). Other details of the chemistry, including the full list of reactions and rate coefficients that we used, are described in HRL96.

For the power-law spectrum in equations (4.4) and (4.7), we choose a power-law index $\alpha = 0.7$, and an upper cutoff of $E = 40$keV for the background flux, as expected from quasar spectra and fits to the X-ray background (Fabian & Barcons 1992). We leave the normalization of the flux as a free parameter, and specify it as J_{21}, the flux (without absorption) at the Lyman limit in units of 10^{-21} erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}.

We calculate the equilibrium fraction of H_2 within a UV-irradiated homogeneous gas cloud by integrating the chemical rate equations. In order to construct the equilibrium H_2 profile as a function of radius, we subdivide the homogeneous cloud into a large number of thin plane-parallel slabs, and in each slab we integrate the rate equations until equilibrium is reached, taking account of the cumulative optical depth due to the H, He, and the H_2 lines. We consider a top-hat overdensity of 200 relative to the background universe, and five different combinations of redshift and temperature. As a plausible example, we first set z=20 and $T = 10^3$K. We then repeat the calculation for z=10 and 50 keeping $T = 10^3$K, and for $T = 10^{2.2}$ and $10^{3.5}$K keeping z=20. We selected these temperatures to bracket the range favorable for H_2 cooling; H_2 is not formed below $10^{2.2}$K, while it is collisionally dissociated above $\sim 10^{3.5}$K (see HRL96). The background universe is assumed to have $\Omega_b = 0.1$ and $H_0 = 50$ km s^{-1} Mpc^{-1}. As both the temperature and the density of the gas are fixed, the only free parameter left in this calculation is the normalization of the flux, J_{21}.

An example for a typical H_2 fraction profile in a cloud is shown in figure 4.2. This figure shows the variation of the equilibrium H_2 fraction f_{H2} as a function of neutral hydrogen column density N_H for a hydrogen density $n_H = 1$ cm^{-3}, a temperature $T = 1000$K, and a flux $J_{21} = 0.1$. Based on the processes which determine the H_2 fraction, three characteristic regimes may be distinguished. Up to hydrogen column densities of $N_H \sim 10^{19}$ cm^{-2}, the most important mechanism for the destruction of H_2 is direct photoionization by photons with energies above 15.4eV. At larger N_H, the destruction rate is dominated by the Solomon process, and f_{H2} drops. At still larger depths, $N_H \sim 10^{23}$ cm^{-2}, the Lyman and Werner lines of H_2 become self-shielding, and collisional destruction by fast electrons (supplied mainly by the ionization of helium) dominates. The equilibrium state deep inside the cloud where $N_H \sim 10^{24}$ cm^{-2} is fully molecular, but this state is not likely to be reached within a Hubble time (see discussion in HRL96).

Figure 4.2 also shows the ratio of the cooling time to the dynamical time t_{cool}/t_{dyn} as a function of depth, where

$$t_{cool} = \frac{(3/2)nk_B T}{\Lambda_{tot}} \qquad \text{and} \qquad t_{dyn} = \frac{1}{\sqrt{6\pi G\rho}}. \tag{4.10}$$

[3]The effect of the sawtooth modulation on the H_2 dissociation rate can be fairly large. In the limit $z_{obs}/z_{on} \to 0$, the modulation decreases the total dissociation rate by a factor of 32.5. This is the maximum effect the sawtooth modulation can introduce, and as z_{obs} approaches z_{on}, the decrease in the rate is smaller [cf. eq. (4.6)].

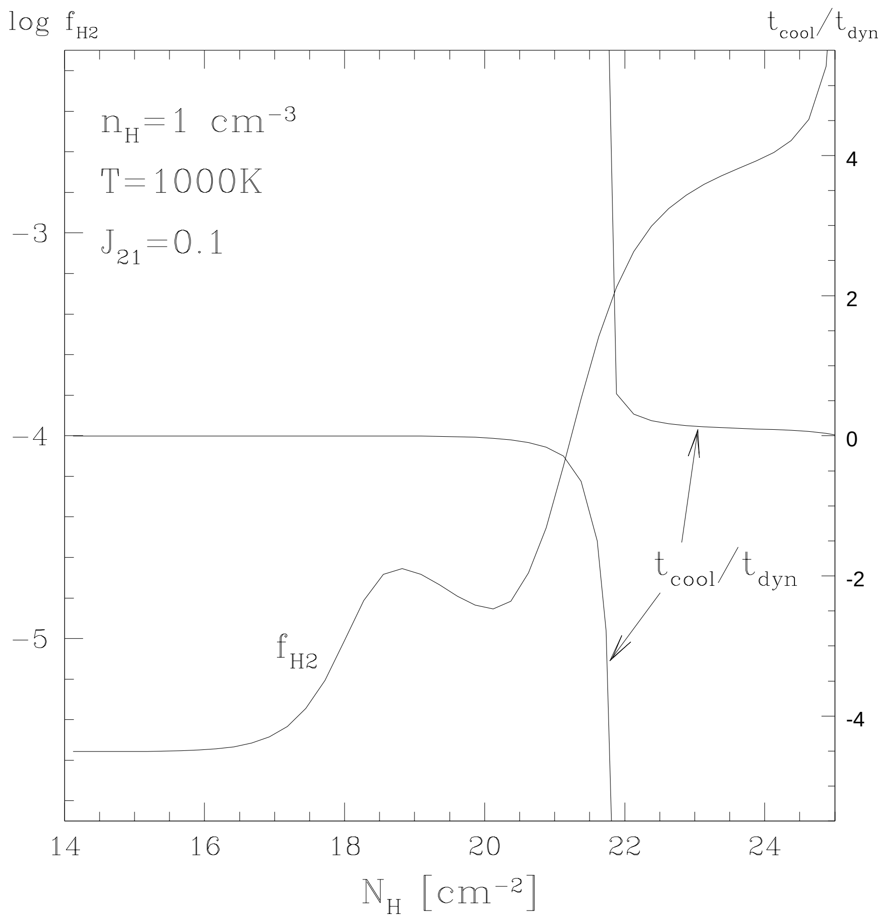


Figure 4.2: *The equilibrium profile of the* H_2 *fraction* f_{H_2}*, and the ratio of the cooling time to the dynamical time* t_{cool}/t_{dyn}*. At low column densities, photoionization results in a net heating, this is expressed by* $t_{cool} < 0$*. At the column density* $N_H \sim 10^{21}$ cm^{-2} *the* H_2 *Lyman and Werner lines become self-shielding and the molecular fraction starts to rise. When the molecular fraction increases above* 10^{-4}*,* H_2 *cooling dominates over photoionization heating, resulting in net cooling* ($t_{cool} > 0$)*. At a slightly larger depth, the* H_2 *cooling time decreases below the dynamical time. The column density at which this occurs,* $N_H \sim 10^{22}$ cm^{-2}*, defines the minimum cooling mass used in § 3.*

The total cooling function Λ_{tot} includes photo-ionization heating, H_2 cooling, and other terms that are summarized in detail in HRL96. The H_2 cooling term was taken from Lepp & Shull 1984. Up to a column density $N_H \approx 10^{22}$ cm^{-2}, photo-ionization heating dominates H_2 cooling, and the total cooling function, as well as t_{cool} are negative, as shown in figure 4.2. At $N_H \approx 10^{22}$ cm^{-2}, the amount of H_2 is sufficiently large so that H_2 cooling is effective, and at this depth t_{cool}/t_{dyn} becomes positive and quickly drops below unity.

Using profiles similar to figure 4.2 for each value of J_{21}, we have found the depth r_{H2} at which a sufficient H_2 abundance is formed so that the H_2 cooling time is shorter than the dynamical time. For each value of J_{21}, we then define the minimum cooling mass to be $M_{H2} \equiv 4\pi/3\rho r_{H2}^3$. Figure 4.3a shows the cooling mass as a function of J_{21} for five different combinations of redshift and temperature. The most probable combination $T = 10^3$ K, $z = 20$ is shown as a thick line. In figure 4.3b we show the same curves, obtained by calculating the minimum cooling mass without the sawtooth

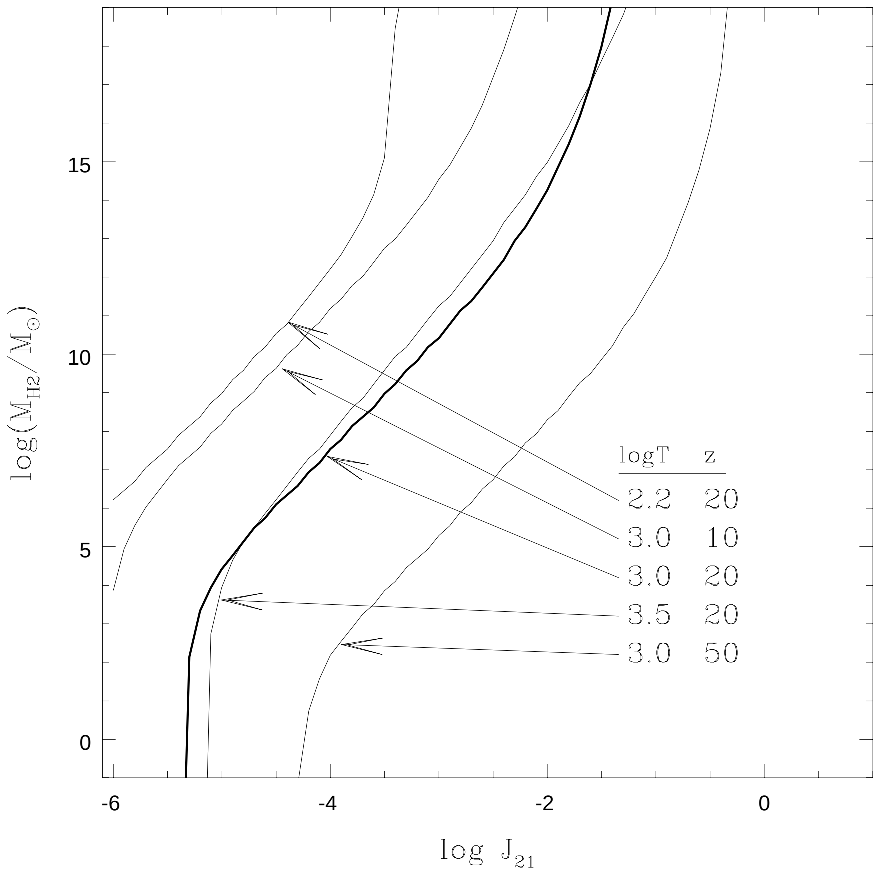


Figure 4.3: *(a) The minimum mass* $M_{\rm H2}$ *of a homogeneous, spherically symmetric cloud which is able to self-shield and form sufficient amount of* $\rm H_2$ *at its center so that it cools on a dynamical time-scale. The lines show* $M_{\rm H2}$ *for five different combinations of redshift and temperature. Among these, the intermediate combination* $(\log T, z) = (3, 20)$ *is shown as a thick line.*

modulation below 13.6eV.

Figures 4.3a and 4.3b bracket the range of possible $\rm H_2$ photodissociation rates, and imply that in general the cooling masses need to be exceedingly large, even when J_{21} has a relatively low value. In particular, the proximity effect confirms the existence of a background flux of at least $J_{21} \approx 0.1$ that keeps the IGM fully ionized at $z \approx 5$. A similar, or even higher value of J_{21} is expected at the time of breakthrough during the reionization epoch at higher redshifts, since the density of the IGM is then higher. Figures 4.3a and 4.3b show that even at fluxes that would be able to ionize the IGM only partially, $J_{21} \lesssim 0.01$, the mass required to shield $\rm H_2$ against the radiation is large, $M_{\rm H2} \gtrsim 10^9 {\rm M}_\odot$. This conclusion can be inferred from both figures 4.3a and 4.3b, implying that it is, in fact, insensitive to the details of the sawtooth-modulation of the spectrum.

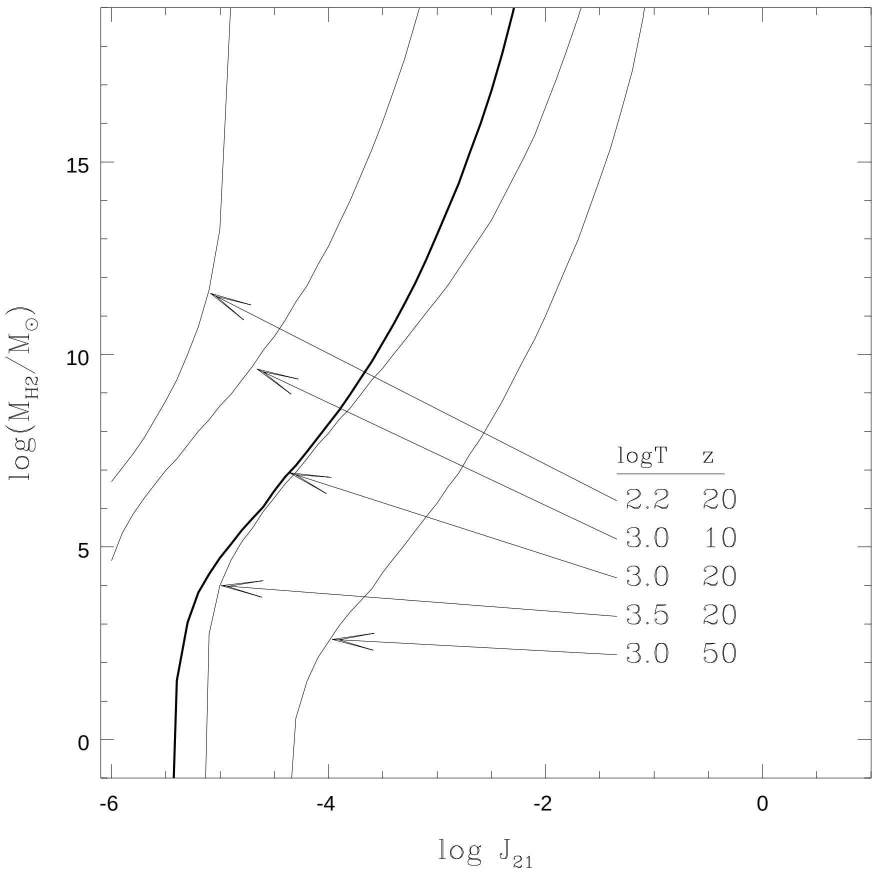


Figure 4.3: *(b) Same as (a), but with the minimum mass calculated without the sawtooth-modulation of the spectrum below 13.6eV. Note that the cooling masses in (a) and (b) differ only slightly. In both cases, they are high already at fluxes that are too low to ionize the IGM.*

4.4 Cosmological Implications

Our main result is that the mass of a virialized primordial cloud has to be exceedingly high in order to self-shield against even a low level of background radiation. Figures 4.3a and b show the mass required to form enough H_2 molecules to cool the cloud on its dynamical time-scale, *versus* the normalization of the flux J_{21}. The minimum cooling masses shown in these figures are indeed high, in all cases they are above $\sim 10^9$ $M_\odot$ already for fluxes that are too low to fully ionize the IGM.

It has been argued in previous works (HTL96; Tegmark et al. 1997) that clouds with virial temperatures 100 K $\lesssim T_{\rm vir} \lesssim 10^4$ K are able to cool by H_2 and continue to collapse after they virialize. These papers, however, did not take into account any external UV radiation. Our present results imply that the appearance of the very first UV sources will suppress the H_2 abundance in these objects, and prevent their post-virialization collapse.

The cooling masses we obtained depend primarily on the number of photons in the narrow spectral range of 11.18eV-13.6eV. Our conclusions are therefore insensitive to the detailed shape of the spectrum, and should hold for either stellar (blackbody) or quasar (power-law) spectra. We also note that in our calculation of the cooling mass, we assumed a homogeneous spherical cloud. Although centrally condensed clouds are, in general, more effective than homogeneous clouds in self-shielding their molecules, a steepening of the density profile is unlikely to decrease the value of the required cooling mass by orders of magnitude.

We have presented our results for a range of temperatures and redshifts. Our results are relevant to any model where the initial fluctuations have amplitudes decreasing with scale, so that cosmic structures form hierarchically. Such models predict different turn-on redshifts for the first sources of radiation; in PIB models, turn-on may occur at $z > 100$; in CDM in the range $z = 10-50$; for mixed dark matter models the first structures may form still more recently. Molecular cooling is more efficient at high densities, and therefore at larger redshifts; however in all these models it determines the scale of the first objects that condense out and contribute the first injection of heat into the universe. The amount of UV generated by each such object is very uncertain; it depends on the efficiency of star formation, and whether the IMF favors massive stars, or even super-massive objects or black holes. However, we have shown that H_2 cooling is quenched even in massive objects before the UV background has reached the level needed to photoionize the entire IGM. Therefore the radiation of the first sources exerts an important feedback on the cosmogonic process: the only clouds able to fragment into new stars are those with virial temperature above $\sim 10^4$K, i.e. with masses $\gtrsim 10^8 \mathrm{M}_\odot[(1+z_{\rm vir})/10]^{-3/2})$, so that isothermal contraction is possible due to atomic line cooling. Clouds that virialize at lower temperatures will, in the absence of a sufficient amount of H_2, maintain virial equilibrium and not fragment into stars.

We therefore draw the robust conclusion that the IGM remained predominantly neutral until a sufficient number of objects with virial temperatures above 10^4 K (masses above $\sim 10^8 \mathrm{M}_\odot[(1+z_{\rm vir})/10]^{-3/2}$) had gone non-linear. Most of the O-B stars or accreting black holes that photoionized the IGM had to form in systems at least as large as this. Formation of such systems would have continued unimpeded until ionization was complete; subsequently J_{21} rises sharply to a value of order unity, when the universe becomes, in effect, a single HII region. The only net cooling of a fully photoionized gas comes from bremsstrahlung, which is less effective than the collisionally-excited line emission from a partially ionized gas (Efstathiou, 1992; Weinberg, Hernquist and Katz 1997, Navarro and Steinmetz 1997). The completion of photoionization may therefore signal another pause in the cosmogonic process, associated with the further decline in the efficiency of cooling, and the consequent further increase in the minimum scale that can collapse.

Chapter 5

Signatures of Stellar Reionization of the Universe

Zoltán Haiman and Abraham Loeb
Astronomy Department, Harvard University, Cambridge, MA 02138

[published in *The Astrophysical Journal*, **483**, 21 (1997)]

ABSTRACT

The high ionization level and non-zero metallicity ($\sim 1\%\ Z_\odot$) of the intergalactic gas at redshifts $z \lesssim 5$ implies that nonlinear structure had started to form in the universe at earlier times than we currently probe. In Cold Dark Matter (CDM) cosmologies, the first generation of baryonic objects emerges at redshifts $z \sim 10$–50. Here we examine the observable consequences of the possibility that an early generation of stars reionized the universe and resulted in the observed metallicity of the Lyα forest. Forthcoming microwave anisotropy experiments will be sensitive to the damping of anisotropies caused by scattering off free electrons from the reionization epoch. For a large range of CDM models with a Scalo stellar mass function, we find that reionization occurs at a redshift $z \gtrsim 10$ and damps the amplitude of anisotropies on angular scales $\lesssim 10°$ by a detectable amount, ~ 10–25%. However, reionization is substantially delayed if the initial stellar mass function transformed most of the baryons into low mass stars. In this case, the mass fraction of pre–galactic stars could be constrained from the statistics of microlensing events in galactic halos or along lines of sight to quasars. Deep infrared imaging with future space telescopes (such as SIRTF or the Next Generation Space Telescope) will be able to detect bright star clusters at $z \gtrsim 5$. The cumulative Bremsstrahlung emission from these star clusters yields a measurable distortion to the spectrum of the microwave background.

5.1 Introduction

Currently, a number of competing models for structure formation in the universe have comparable success in accounting for the observed large scale structure at $z \lesssim 5$ (based on galaxy surveys and quasar absorption line studies) and at $z \sim 10^3$ (based on microwave anisotropy experiments). Thus, an efficient method for identifying the unique model that describes structure formation in the real universe is to study the early collapse of small mass objects at the intermediate redshift interval $5 \lesssim z \lesssim 10^3$. Variants on a popular cosmology which make similar predictions about the formation of structure on large scales, might differ *qualitatively* in their predictions about structure on small scales. For example, the standard Cold Dark Matter (CDM) power–spectrum of fluctuations predicts a bottom–up hierarchy of structure formation which is only mildly (logarithmically) satisfied at the low mass end of the fluctuation spectrum. Any power–law tilt of this spectrum that adds more power on large scales would change the hierarchy of low–mass objects to that of a top–down scenario, reminiscent of hot dark matter.

In CDM cosmologies, the first baryonic objects with masses $\sim 10^5 M_\odot$ form at a redshift as high as $z \sim 50$ (see Haiman, Thoul & Loeb 1996, hereafter HTL96, and references therein). It has long been realized that star formation in these objects could release a sufficient flux of UV photons to reionize the universe (Couchman & Rees 1986; Fukugita & Kawasaki 1994; Shapiro, Giroux, & Babul 1994; Tegmark, Silk & Blanchard 1995; Tegmark et al. 1996; Ostriker & Gnedin 1996). Since the ionization potential of hydrogen (13.6 eV) is 5–6 orders of magnitude smaller than the nuclear energy release per baryon in a star (6.6 MeV), only a negligible fraction of all the baryons need to be processed through stars in order for the universe to get reionized. An even smaller fraction is necessary if the progenitors of quasar black holes form at the same time (Eisenstein & Loeb 1995).

The reionization epoch begins as each ionizing source blows an expanding Strömgren sphere around itself, and ends when these HII regions overlap (Arons & Wingert 1972). However, long before the Strömgren spheres overlap and breakthrough occurs, molecular hydrogen gets universally dissociated by photons with energies in the range 11–13.6 eV, to which the neutral intergalactic gas is transparent. As a result, molecular cooling is suppressed even inside dense objects (Haiman, Rees, & Loeb 1997, hereafter HRL97). Since efficient cooling is required for stars to form through fragmentation of the metal–poor primordial gas, the abundance of stars builds up to levels capable of ionizing the universe only after massive objects ($\gtrsim 10^8 M_\odot$) form, inside which the high virial temperature of $\gtrsim 10^4$K allows fragmentation to occur through atomic line cooling.

Spectral observations of quasars at $z \lesssim 5$ show evidence for an early reionization epoch. The lack of complete absorption shortward of the Lyα resonance (the so–called Gunn–Peterson effect; Gunn & Peterson 1965) implies that the universe is highly photoionized by $z \sim 5$. The source of the ionizing UV background is unknown, even at relatively low redshifts. Several authors have noted that radiation from quasars alone might be insufficient to reionize the universe (Shapiro & Giroux 1987; Miralda-Escudé & Ostriker 1990), and recent observations of HeII absorption favor a relatively soft spectrum of the ionizing radiation, which is more typical of stars than of quasars (Vikhlinin 1995; Haardt & Madau 1996; but see Davidsen, Kriss & Zheng 1996 for a different view).

The expected reionization of the universe at $10 \lesssim z \lesssim 10^2$ has a number of observable consequences (see the early overview by Carr, Bond & Arnett 1984). First, the resulting optical depth of the universe to electron scattering, $\tau_{\rm es}$, damps the microwave background anisotropies on scales $\lesssim 10°$ (Efstathiou & Bond 1987; Kamionkowski, Spergel, & Sugiyama 1994; Hu & White 1996). The recently detected hint for a Doppler peak in ground–based microwave anisotropy experiments can already be used to set the constraint $\tau_{\rm es} \lesssim 1$ (Scott, Silk & White 1995; Bond 1995).

Future satellite experiments (such as MAP or PLANCK[1]) will be able to probe or limit values of $\tau_{\rm es}$ as small as ~ 0.1. Another consequence of the early epoch of star formation is the metal enrichment of the intergalactic medium (IGM). Early star formation provides a likely explanation of the metal abundance ($\sim 1\% Z_{\odot}$) which is detected at $z \sim 3$ in Lyα absorption systems with vastly different column densities (Cowie et al. 1995; Tytler et al. 1995; Songaila & Cowie 1996; Cowie 1996). Although the level of enrichment in different systems varies by an order of magnitude, the ubiquitous existence of heavy elements in systems with low HI column densities $\mathrm{N_{HI}} \gtrsim 10^{15}$ cm^{-2} is naturally explained by a pre-enrichment phase (Songaila & Cowie 1996). Finally, an old population of pre-galactic stars would behave as collisionless matter during galaxy formation and populate the diffuse halos of galaxies. Such stars might explain some of the events observed by ongoing microlensing searches in the halo of the Milky Way galaxy (see, e.g. Alcock et al. 1996 and references therein; Flynn, Gould & Bahcall 1996) and could also be detected through their lensing effect along lines of sight to distant quasars (Gould 1995).

In this paper, we quantify the observational signatures of the first population of stars in CDM cosmologies. We follow the formation of star–forming regions out of primordial density fluctuations according to the Press–Schechter theory, and calibrate the fraction of gas which is converted into stars based on the inferred carbon abundance in Lyα absorption systems. Our semi–analytic approach incorporates a variety of subtle gas dynamical processes, including the effect of baryonic pressure on the abundance and sizes of virialized regions, the significance of the initial mass function (IMF), the UV spectra and metallicity yields of the stars, the negative feedback of star–formation due to the photodissociation of molecular hydrogen, the absorption of ionizing photons within the dense parent star–forming clouds, and the temporal evolution of the ionization fronts around star clusters. Previous discussions in the literature have discussed some, but not all, of these processes in full detail.

Since some of the details about star formation and radiative transfer could *control* the solution to the reionization problem, a semi–analytic approach appears to be best suited for sorting out the significance of each of these ingredients under different model assumptions. Numerical simulations of the type described by Ostriker & Gnedin (1996) are expensive in computer time, and are still limited by the approximate treatments of star formation and radiative transfer. In our calculation, we follow the evolution of the metallicity of the intergalactic gas, the volume filling factor of ionized regions, and the optical depths of the universe to electron scattering and microlensing, over a large range of redshifts and mass scales. In contrast with numerical simulations, we are able to calculate only the universal average of each quantity; however, the versatility of our semi-analytic approach allows us to examine the sensitivity of our results to many model parameters, such as the power spectrum of primordial density fluctuations, the stellar IMF and metallicity yields, the baryon density parameter $\Omega_{\rm b}$, the escape fraction of ionizing photons from the parent star clusters, and the presence of a negative feedback on star formation due to the photo–dissociation of molecular hydrogen.

The paper is organized as follows. In § 2 we provide a detailed description of all the ingredients of our model. In § 3 we describe how the various observable quantities (metallicity, optical depth to electron scattering, and microlensing probability) are calculated from the model. In § 4 we present our results and examine their sensitivity to model parameters. Finally, § 5 summarizes the main conclusions and implications of this work.

[1] See the homepages for the MAP (http://map.gsfc.nasa.gov) and the PLANCK (http://astro.estec.esa.nl/SA-general/Projects/Cobras/cobras.html) experiments.

5.2 Description of the Model

In contrast with the sudden nature of standard recombination at $z \approx 1400 \pm 100$, reionization is a continuous phase transition that lasts for a considerable fraction of its Hubble time. Initially, most of the intergalactic medium (IGM) is neutral, and a number of ionizing sources of different luminosities and lifetimes turn on. Each of these sources creates an ionized bubble which expands into the IGM at a rate dictated by the source luminosity and the background IGM density. Reionization is completed when these bubbles overlap to fill the entire volume (see Arons & Wingert 1972 for an early discussion of this process in relation to quasars).

Energetically, it is sufficient to convert a small fraction of all the baryons into stars ($\sim 10^{-3}$, see §2.3) in order to ionize the IGM. This implies that the universe could have been already ionized when only a small fraction of its mass was incorporated into nonlinear objects. Therefore, at this stage, each of the rare star–forming objects collapses in isolation and generates an HII bubble which expands into a relatively smooth intergalactic medium. Since the lifetime of UV–bright stars is much shorter than the Hubble time at $z \lesssim 10^2$, the volume filling fraction of HII grows primarily due to the rapid increase in the collapsed baryonic fraction with redshift. The initial growth of this fraction is exponential for a Gaussian random field of density fluctuations.

The abundance and mass distribution of virialized dark matter halos is described by the Press–Schechter theory (Press & Schechter 1974). Due to their finite initial pressure, the baryons are unable to accrete onto shallow potential wells. In our calculations, we take into account the pressure of the baryons and employ spherically symmetric simulations to calculate the abundance and structure of baryonic objects with low masses (for a complete description of these simulations, see HTL96).

As a result of the virialization process, the gas inside each collapsed object settles into a smooth pressure–supported configuration. Further collapse and fragmentation of the gas into stars is possible only as a result of additional cooling. Objects with masses near the Jeans mass ($\sim 10^5 M_\odot$), are characterized by a virial temperature $\sim$ 100–300K. At these temperatures, only cooling through the vibrational and rotational transitions of molecular hydrogen (H_2) is efficient on a dynamical time–scale in a gas of a primordial composition (HTL96). However, the trace amounts of H_2 can be easily photodissociated by a relatively weak flux of photons with energies between 11–13.6 eV, to which the universe is transparent even before reionization (HRL97). The resulting negative feedback that suppresses fragmentation into stars due to the photo–dissociation of H_2 molecules, occurs shortly after the very first ionizing sources appear and prevents further star formation until objects with virial temperatures $\sim 10^4$K collapse, inside which atomic line cooling allows fragmentation (HRL97). Therefore, we assume in our standard model that star formation occurs only inside objects with virial temperatures $\gtrsim 10^4$K. These objects are also immune to photo-ionization heating and could therefore retain their gas even after the universe gets reionized.

The fraction of gas converted into pre–galactic stars determines the average metallicity of the IGM. While the local value of the metal enrichment might vary from place to place due to inefficient mixing, its line-of-sight average over cosmological scales should admit a universal value. This expectation is confirmed by the detection of a roughly (to within an order of magnitude, cf. Rauch, Haehnelt & Steinmetz 1996; Hellsten et al. 1997) universal C/H ratio in Lyα absorbers over a wide range of column densities at $z \sim$ 2–4 (Cowie et al. 1995; Tytler et al. 1995; Songaila &

Cowie 1996, Cowie 1996). The detected carbon might have been expelled out of an abundant population of star forming regions with shallow potential wells, through winds driven by supernovae or photoionization heating.

For a given mass function of stars, the inferred metallicity of the IGM can be used to fix the overall fraction of gas which was converted into stars. In our model, we assume that this fraction is the same for all gas clouds and arrange the total carbon output from all stars to match the observed C/H ratio of $\sim 1\%$ solar at a redshift $z = 3$. We further postulate that star formation occurs in a starburst with a universal Scalo (1986) initial mass function (IMF). The consequences of varying these model assumptions are described in §4.

The stellar IMF determines the composite spectrum of ionizing radiation which emerges from the star clusters. We follow the time-dependent spectrum of a star of a given mass based on the Kurucz (1993) spectral atlases and the evolution of the star on the H–R diagram as prescribed by the tracks in Schaller et al. (1992). The ionizing photons which are emitted from the stellar surfaces can still be consumed by recombinations inside their parent gas cloud before they reach the IGM. In modeling the escape fraction of UV photons from a parent star–forming cloud, we adopt the equilibrium density profile of gas inside each cloud according to our spherically–symmetric simulations (HTL96).

Finally, the ionizing photons which escape each cloud generate an HII bubble in the surrounding IGM. We use the time–dependent luminosity of each star–forming region in calculating the propagation of this ionization front. Since the average separation between stellar sources at the time of reionization is much shorter than the Hubble length, the propagation of photons from a given source to its ionization front can be approximated as instantaneous. In our discussion, we ignore the possible contribution from early quasars to the ionizing background (Eisenstein & Loeb 1995). Our account of the HII bubble dynamics around each time–dependent source differs from several recent discussions which simply calculated the ionization fraction of the universe based on the total energy input and the chemistry of a uniform IGM (e.g. Meiksin & Madau 1993, Fukugita & Kawasaki 1994, Giroux and Shapiro 1996)

5.2.1 The Number Density of Ionizing Sources

The Press–Schechter theory relates the comoving number density, $N_{\rm ps}(z,M)dM$, of collapsed objects with total masses between M and $M + dM$ at a redshift z, to the power spectrum of density fluctuations $P(k)$ (see, e.g. Padmanabhan 1993). In this paper we adopt the BBKS power spectrum for CDM cosmologies (Bardeen et al. 1986), which is parameterized by the primordial power spectrum index n and normalization $\sigma_{8h^{-1}}$. We also assume throughout the paper a cosmological density parameter of $\Omega_0 = \Omega_{\rm b} + \Omega_{\rm dm} = 1$ and a Hubble constant of $h \equiv H_0/100$ km s^{-1} Mpc$^{-1} = 0.5$.

In situations where gas pressure is negligible, the Press–Schechter theory describes the behavior of the baryons as well. The comoving number density of objects with a baryonic mass between $M_{\rm b}$ and $M_{\rm b} + dM_{\rm b}$ is then given by $N_{\rm b}(z, M_{\rm b}) = N_{\rm ps}(z, M_{\rm b}\Omega_0/\Omega_{\rm b})$, where $\Omega_{\rm b}$ is the cosmological density parameter of the baryons. When incorporating the effect of gas pressure, it is often assumed that no baryonic objects form below the Jeans mass predicted by linear–theory $M_{\rm Jeans} \sim 10^5 M_\odot$, and so $N_{\rm b}(z, M < M_{\rm Jeans}) = 0$. In reality, this assertion is inaccurate as the gas pressure merely delays the collapse of low–mass baryonic systems but does not prevent it. Since the history of reionization is sensitive to trace amounts of collapsed gas, it

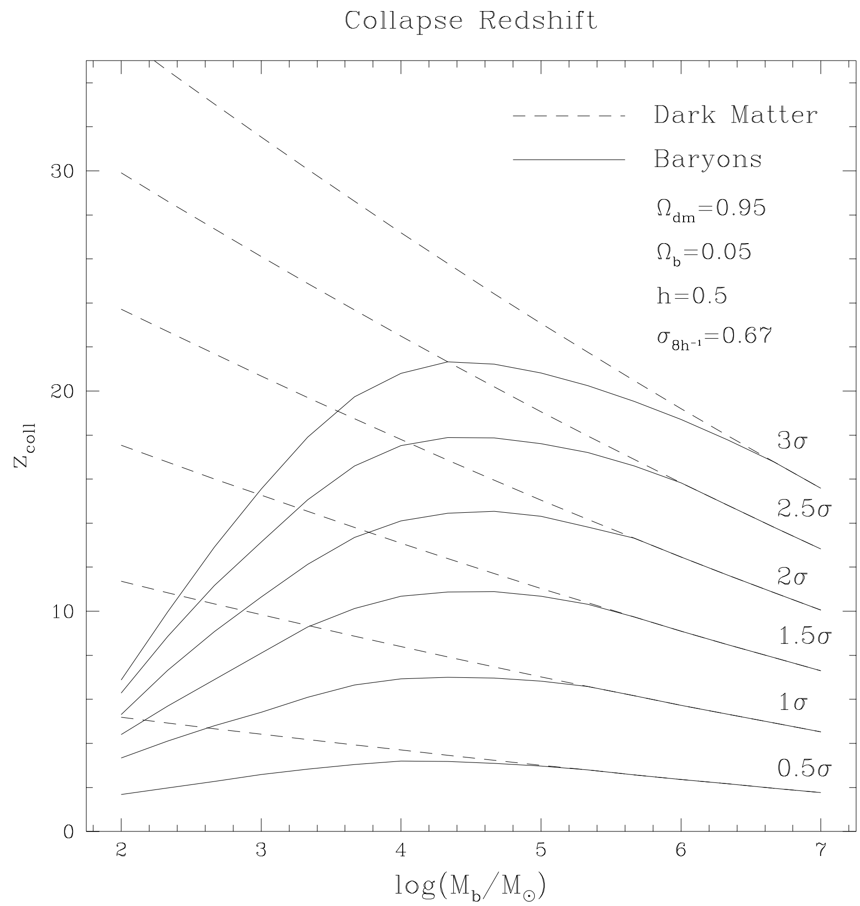


Figure 5.1: *Collapse redshift for cold dark matter (dashed lines) and baryons (solid lines) in spheres of various masses and overdensities (the latter in units of $\sigma(M)$ for a BBKS power spectrum). The collapse of the baryons is delayed relative to the dark matter due to gas pressure. The curves were obtained by following the motion of the baryonic and dark matter shells with a spherically symmetric, Lagrangian hydrodynamics code (HTL96). The dashed lines agree with the collapse redshift for the dark matter based on the linear extrapolation of the spherical collapse model for pressureless matter, $z_{\rm coll} = \nu\sigma/1.69 - 1$.*

is important to include a more accurate treatment of the gas dynamics in low mass objects which are first to produce ionizing photons. We incorporate the effect of gas pressure into $N_{\rm b}(z, M_{\rm b})$ based on runs from our simulations of spherically symmetric clouds (HTL96). Our approach is the simplest extension of the spherical top–hat collapse model which proved to be successful when used to define the linear overdensity threshold for the collapse of objects in the Press–Schechter mass function. While the top–hat model is adequate for defining the characteristic collapse time of the dark matter, it must be modified for the baryons due to their pressure.

Our simulations start with a slightly overdense sphere shortly after recombination. The initial conditions are chosen to represent spherically symmetric peaks of various widths and heights in the linear overdensity field at $z = 10^3$. If the mass of the sphere is well above the Jeans mass, pressure can be ignored, and the baryonic component expands, turns around, and collapses together with the dark matter component. However, for an object close or below the Jeans mass, gas pressure delays

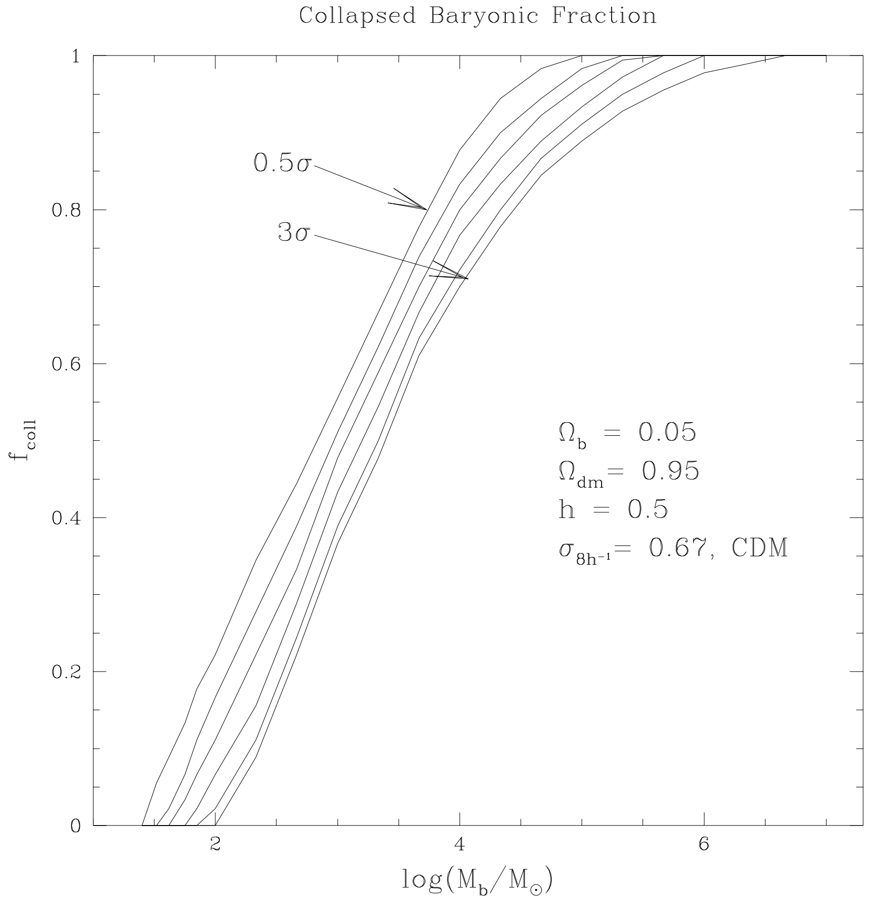


Figure 5.2: *The fraction of baryons that managed to collapse by the time the outermost shell of the dark matter halo had collapsed. The curves were obtained with the same hydrodynamics code as in Figure 5.1. Six different curves are shown for the six different peak heights,* $0.5\sigma - 3\sigma$*, as in Figure 5.1. These curves yield an "equivalent cosmological Jeans mass" of* $10^3 M_\odot$ *(cf. § 2.1).*

the collapse of the baryons relative to the dark matter. The smaller the mass of the system is, the longer the delay gets. We obtained the exact collapse redshifts by following the motion of both the baryonic and the dark matter shells with our code. Figure 5.1 shows the collapse redshift of both components for spheres of different initial masses and overdensities. This figure demonstrates that objects with baryonic masses as low as $\sim 10^2 \mathrm{M}_\odot$ (i.e., well below the Jeans mass predicted by linear theory) are still expected to form before $z \sim 20$.

In order to incorporate the results shown in Figure 5.1 into $N_b(z, M_b)$, we determined in each run the fraction of the baryonic mass which managed to collapse before the virialization of the outermost dark matter shell. Figure 5.2 shows this fraction as a function of the total baryonic mass for various initial overdensities. Based on the collapse fraction f_{coll}, the number density of objects is given by $N_b(z, M_b) = N_{ps}(z, M_b\Omega_0/f_{coll}\Omega_b)$.

The total fraction F_{coll} of baryons contained within collapsed objects at a redshift z can then be expressed in terms of f_{coll} and N_{ps},

$$F_{coll}(z) = \frac{1}{\rho_0}\int_0^\infty dM f_{coll}(z, M) M N_{ps}(z, M), \tag{5.1}$$

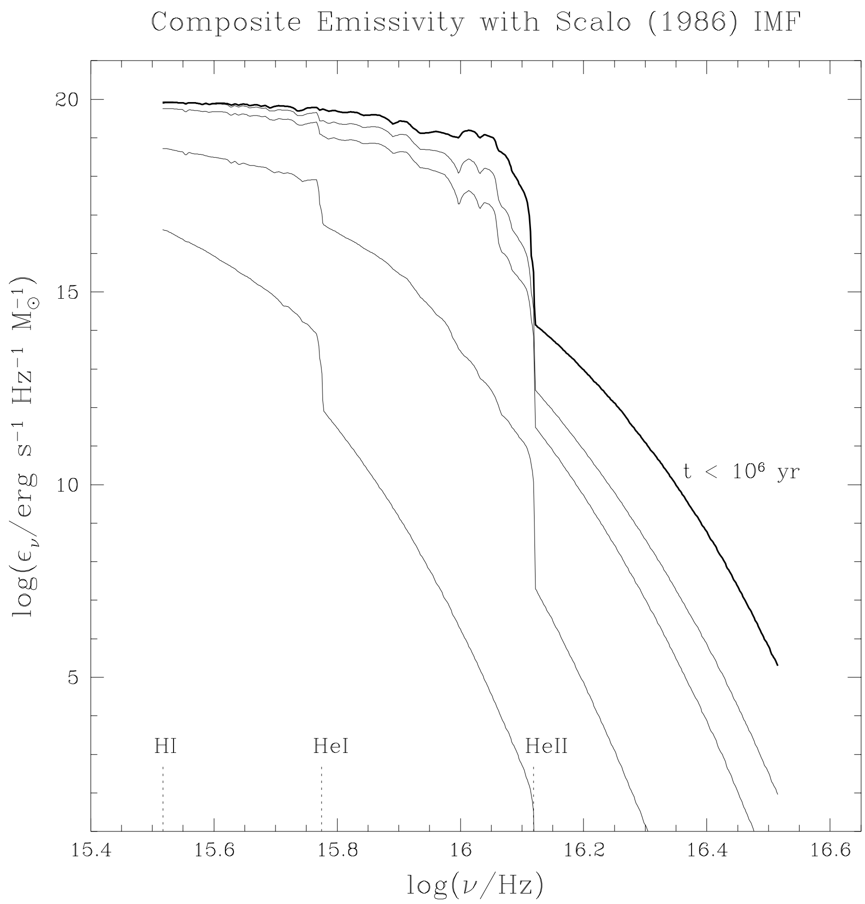


Figure 5.3: *Composite emissivity per solar mass for a star cluster with a Scalo IMF. Each star is assumed to evolve on the H–R diagram according to the theoretical evolutionary tracks of Schaller et al. (1992) and to emit radiation according to the spectral atlas of Kurucz (1993). The thick line shows the steady emissivity when all the stars are still on the main sequence (* $\lesssim 10^6$ *years after the starburst). The four solid lines show the emissivity at four subsequent times (from top to bottom): t=3, 4, 10, and 30 million years after the starburst. Frequencies corresponding to the ionizations of hydrogen and helium are marked – note the sharp drop of the emissivity at the HeII edge.*

where ρ_0 is the mean present-day density of the universe, and $f_{\rm coll}$ is a function of the collapse redshift and mass of the dark matter halo. It is interesting to define an "equivalent Jeans mass" by approximating $f_{\rm coll}$ as a step function with the step located at the mass scale $M_{\rm crit}$ that provides the best fit to the exact $F_{\rm coll}(z)$, obtained with the full $f_{\rm coll}(z, M)$. Using this definition, we obtain a critical mass $M_{\rm crit} \approx 10^3 M_\odot$ before reionization, which is about 1–2 orders of magnitude smaller than the commonly adopted value for the cosmological Jeans mass.

5.2.2 Internal Structure of Ionizing Sources

Our model assumes that a constant fraction $f_{\rm star}$ of the gas inside each object turns into stars in an instantaneous starburst with a Scalo (1986) IMF. Since the IMF could be different at high redshifts than it is observed locally, we also consider a tilted Scalo

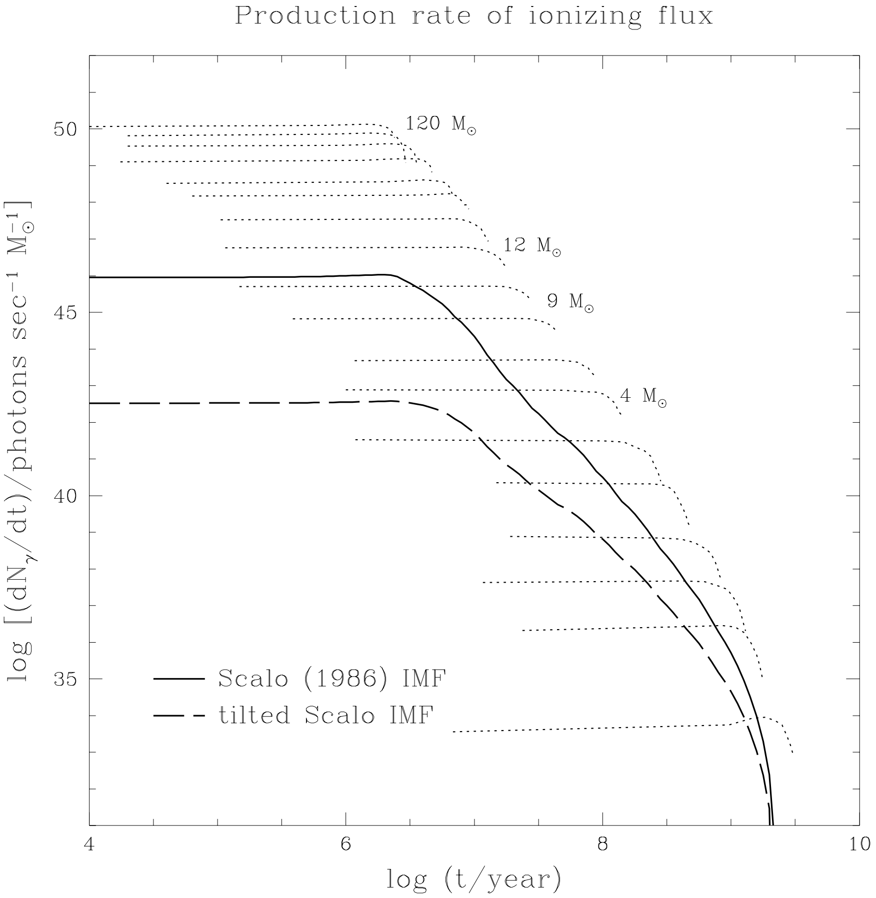


Figure 5.4: *The boldface solid line shows the number of ionizing photons emitted per second per solar mass (without the factors $f_{\rm esc}$ and $f_{\rm star}$, see § 2.3) for the composite emissivity shown in Figure 5.3 and a Scalo IMF. The boldface dashed line shows the rate for a tilted Scalo IMF, where most of the stars have low masses. The dotted lines show the separate contributions from stars with masses (top to bottom) 120, 85, 60, 40, 25, 20, 15, 12, 9, 7, 5, 4, 3, 2.5, 2, 1.7, 1.5, and 1.25 $M_\odot$, in units of photons/sec.*

function which is defined by adding a constant β to the power law indices of the Scalo IMF while keeping the IMF normalization fixed at $4\mathrm{M}_\odot$. This tilt leaves the total carbon yield almost unaffected but changes the net fraction of gas which is converted into stars (see discussion below).

For a star of a given mass, we find the time evolution of the effective temperature ($T_{\rm eff}$) and surface gravity ($g_{\rm surf}$) from the low metallicity ($Z = 0.001$) evolutionary tracks in Schaller et al. 1992. For each pair of $T_{\rm eff}$ and $g_{\rm surf}$ values, we then obtain the flux of the star $F(\nu)$ using the spectral atlas of Kurucz (1993) at the lowest metallicity available, $Z = 10^{-5}$. The resulting time–dependent composite emissivity ϵ_ν per solar mass is shown in Figure 5.3. The top curve shows the initial emissivity, which remains steady over $\sim 10^6$ years after the starburst. Subsequently, the emissivity drops as stars with progressively lower masses leave the main sequence.

In this paper we ignore the emission of ionizing radiation from supernova explosions. Assuming that each star more massive than $8M_\odot$ explodes with a hydrodynamic energy output of 4×10^{50} erg (Woosley & Weaver 1986), the supernova energy release per stellar mass, averaged over a Scalo IMF, is 1.2×10^{48} erg $M_\odot^{-1}$, i.e. only $\sim 3\%$ of the energy released in stellar emission of ionizing photons from the same IMF, 4.4×10^{49} erg $M_{\odot}^{-1}$.

Figure 5.4 shows the expected evolution in the production rate of ionizing photons for stars with various masses between 1 and $120 M_{\odot}$ (dotted curves) and for the composite flux of a Scalo IMF (solid curve). The ionizing flux starts to decline roughly two million years after the starburst due to the limited lifetime of stars on the main sequence. The total number of ionizing photons which are released per proton processed through a star is ~ 1100, about two order of magnitudes lower than the corresponding value for a single O star. The dashed line shows the production rate of ionizing photons when the standard IMF is tilted by a power–law index of β=1.69 in the direction of introducing more low mass stars; this choice of β is motivated by the possibility that the halos of galaxies are composed of Massive Compact Halo Objects (MACHOs). The lack of massive stars in the tilted IMF reduces significantly the number of ionizing photons to a total value of 0.66 ionizing photons per baryon processed through a star.

We assume that fragmentation converts a constant fraction of the gas into stars at every point inside the cloud. However, not all of the ionizing photons which are emitted by these stars, escape the cloud and ionize the IGM. The escape probability $f_{\rm esc}$ depends on the density profiles of both the gas and the stars inside the cloud. For both components, we adopt a radial density profile that varies as $\propto 1/r^2$ and is truncated at some outer radius, where the density jumps by a factor of ~ 50 from its background value. This profile is in good agreement with the numerical results described in HTL96, and is slightly shallower than the dark matter profile ($\propto 1/r^{2.25}$) predicted by self–similar solutions (Bertschinger 1985). Under the assumption of a constant $f_{\rm star}$ throughout the cloud, we calculate the equilibrium ionization profile both inside and outside the cloud by solving the equations of radiative transfer and ionization equilibrium,

$$F_\nu(r) = \frac{1}{r^2}\int_0^r r'^2 e^{-\tau_\nu(r,r')} j_\nu(r')dr', \tag{5.2}$$

$$\tau_\nu(r,r') = \sigma_\nu\left[N_{\rm H}(r) - N_{\rm H}(r')\right], \tag{5.3}$$

$$\alpha_{\rm B}(T) n_{\rm e}^2 = n_{\rm H}\int_{\nu_{\rm H}}^{\infty} \frac{4\pi\sigma_\nu F_\nu}{h_{\rm P}\nu}d\nu, \tag{5.4}$$

where $j_\nu = \epsilon_\nu f_{\rm star}\rho_{\rm b}(r)/(4\pi)$ is the monochromatic emission coefficient in units of erg cm^{-3} s^{-1} ster^{-1} Hz^{-1}; F_ν is the flux in units of cm^{-2} s^{-1}; $\tau_\nu(r,r')$ is the optical depth between radii r and r'; σ_ν is the ionization cross–section of hydrogen in units of cm^2; $N_{\rm H}(r)=\int_o^r n_{\rm H}(r)dr$ is the column density of neutral hydrogen between the center and a radius r; $\alpha_{\rm B}(T)$=2.6×10^{-13} cm^3 s^{-1} is the case B recombination coefficient at $T \approx 10^4$K; $n_{\rm e}$ and $n_{\rm H}$ are the number densities of electrons and neutral atoms in cm^{-3}; $\rho_{\rm b}(r)$ is the mass density profile of the gas; $\nu_{\rm H}$ is the ionization threshold of hydrogen, and $h_{\rm P}$ is Planck's constant.

Figure 5.5 shows an example of the resulting ionization profile for an object with a baryonic mass of $10^6 M_{\odot}$ at $z = 50$. The size of this object is ~ 150 pc, and the vertical dashed line at ~ 2.5 kpc shows the size that the associated Strömgren sphere would have had if the recombinations inside the object were ignored and only the recombinations in the background gas were taken into account. The recombinations inside the object reduce the size of the Strömgren sphere to ~ 0.8kpc, from where we infer that only a fraction $f_{\rm esc} \approx (0.8/2.5)^3 = 3.3\%$ of the ionizing photons escaped from the object.

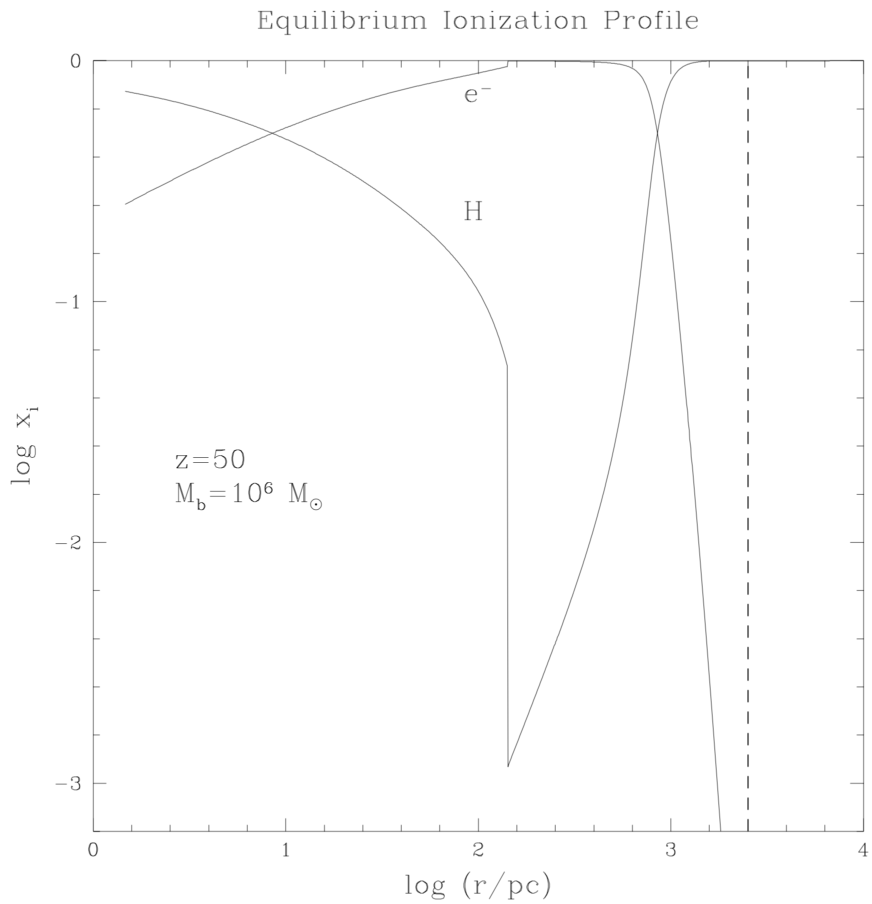


Figure 5.5: *Example of an equilibrium ionization profile inside and around a spherically symmetric object: the fractions* $x_{\rm e} = n_{\rm e}/(n_{\rm e} + n_{\rm H})$ *and* $x_{\rm H} = n_{\rm H}/(n_{\rm e} + n_{\rm H})$ *are shown as a function of radius. We assume that the object has a baryonic mass of* $10^6 M_\odot$ *and a density profile* $\propto 1/r^2$*, with a factor of 50 drop in the density at the edge of the object* ($r \sim 150pc$)*, outside of which the density is uniform and equal to the background value at z=50. A fraction* $f_{\rm star} = 4\%$ *of the baryons in the object are converted into stars with the emission spectrum shown in Figure 5.3. The vertical dashed line shows the radius of the Strömgren sphere when recombinations in the object are ignored; this can be used to infer the escape fraction of ionizing photons of 3.3%.*

We have computed similar equilibrium ionization profiles for objects of different masses and redshifts, and in each case we obtained the escape fraction $f_{\rm esc}$. The results are shown in Figure 5.6 for objects with masses between $10^4 M_\odot$ (top curve) and $10^9 M_\odot$ (bottom curve). This figure demonstrates that the escape fraction varies only slightly with mass but depends strongly on redshift. At redshifts below $z = 10$ the escape fraction is close to unity; at higher redshifts one can derive a useful fitting formula, accurate to within $\sim 10\%$,

$$\log f_{\rm esc} = \begin{cases} 1.92\{\exp[-(z-10)^2/1510] - 1\} & (\text{for } z > 10) \\ 0 & (\text{for } z \leq 10). \end{cases} \tag{5.5}$$

Although the escape fraction of the ionizing photons is a crucial ingredient in all of our reionization models, little is known about its value in nearby star forming

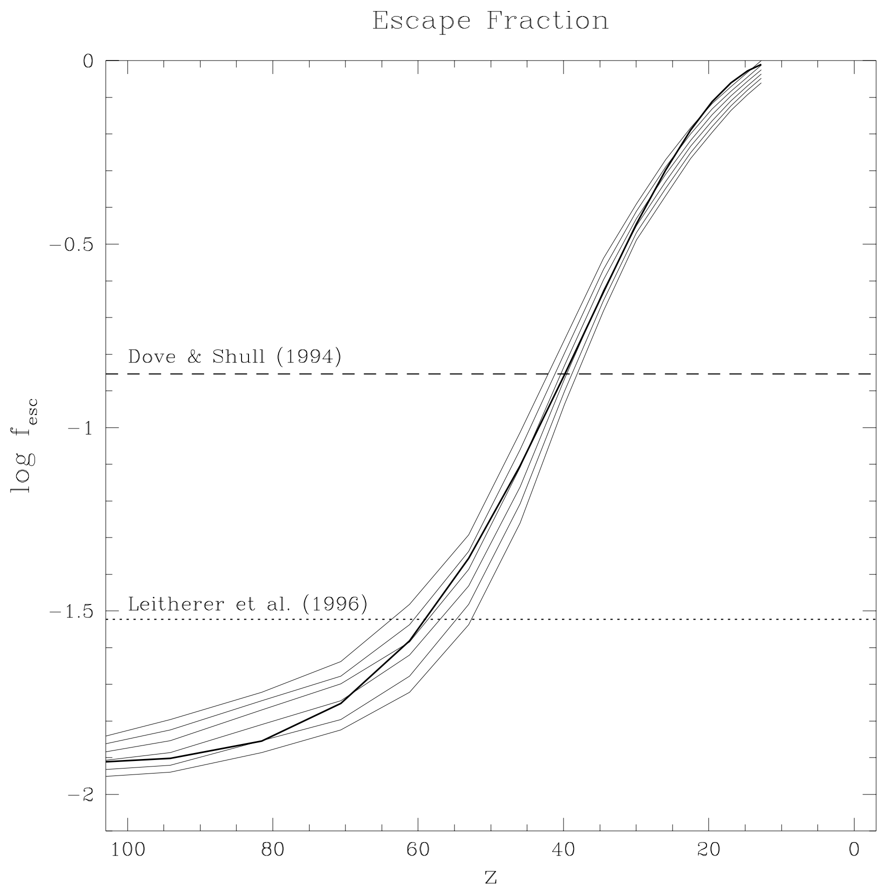


Figure 5.6: *Escape fraction of ionizing photons from objects of different masses at various redshifts. The curves from bottom to top correspond to baryonic masses* $10^{4,5,6,7,8,9}$ $M_\odot$*, and are obtained from equilibrium ionization profiles similar to Figure 5.5. The boldface solid line shows the fitting formula* $\log f_{\rm esc} = 1.92(\exp[-(z-10)^2/1510]-1)$ *for* $z > 10$*. For comparison, the dashed line shows the theoretical estimate of the escape fraction from the disk of the Milky Way galaxy (Dove & Shull 1994), and the dotted line shows the observed value of the escape fraction in four nearby starburst galaxies (Leitherer et al. 1996).*

regions. For reference, it is possible to compare our estimates with the value of 14% obtained by Dove & Shull (1994) in their theoretical modeling of HII regions around OB associations inside the exponential disk of the Milky Way galaxy (short dashed line in Fig. 5.6). Another possible comparison is with the value of 3% reported by Leitherer et al. (1995) from observations of four nearby starburst galaxies (long dashed line). On the one hand, our approach overestimates the escape fraction because it ignores small scale clumpiness of the gas. On the other hand, it underestimates the escape fraction, because the system might lose its gas due to winds driven by supernova explosions or photoionization heating. The dependence of our results on changes in the adopted values of the escape probability is discussed in §4.

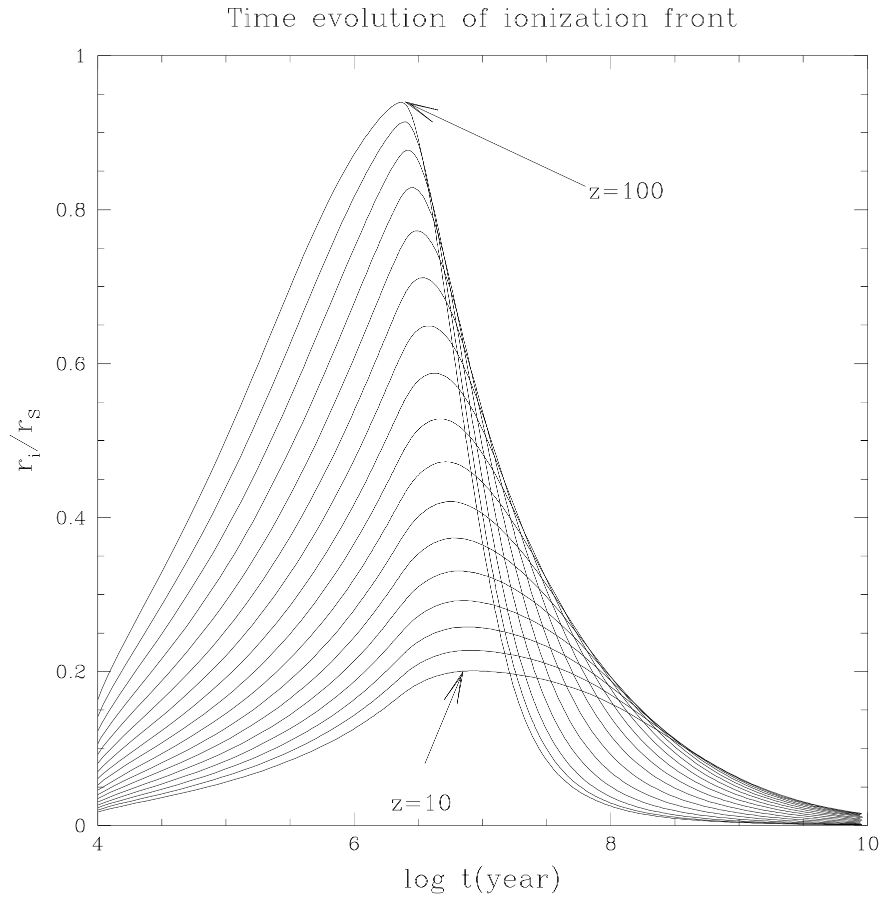


Figure 5.7: *Radius of the ionization front in units of the time-varying Strömgren radius (cf. eq. 5.8), for sources that turn on between redshifts z=10 and z=100, and have the photon production rate shown by the solid line in Figure 5.4. The turn-on redshifts of the 17 curves are equally spaced in log z.*

5.2.3 Propagation of the Ionization Fronts

The propagation of ionization fronts in an expanding medium was considered by Shapiro & Giroux (1987) in the context of reionization of the universe by quasars. The temporal evolution of the radius of the ionization front $r_{\rm i}$ in physical (i.e. not comoving) coordinates is determined by the equation

$$n_{\rm H}\left(\frac{dr_{\rm i}}{dt}-H(t)r_{\rm i}\right)=\frac{1}{4\pi r_{\rm i}^2}\left(\frac{dN_\gamma}{dt}(t)-\frac{4}{3}\pi r_{\rm i}^3\alpha_{\rm B}n_{\rm H}^2\right), \tag{5.6}$$

where $H(t)$ is the Hubble constant, $dN_\gamma/dt(t)$ is the emission rate of ionizing photons by the source, and $n_{\rm H}$ is the background number density of neutral hydrogen in the IGM. Equation (5.6) simply expresses the fact that during the expansion of the ionization front, the ionization rate of fresh atoms is equal to the difference between the emission rate of photons from the source and the total recombination rate in the already ionized region. In the case of a steady source (dN_γ/dt=const), equation (5.6) can be solved analytically (Shapiro & Giroux 1987) for $r_{\rm i}(t)$; with a time-dependent source, we substitute $dN_\gamma/dt(t)$ and obtain $r_{\rm i}(t)$ numerically. In our model,

$$\frac{dN_\gamma}{dt}(t) = f_{\rm esc} f_{\rm star} M_{\rm b} \int_{\nu_{\rm H}}^{\infty} d\nu \frac{\epsilon_\nu}{h_{\rm P}\nu}, \tag{5.7}$$

where $M_{\rm b}$ is baryonic the mass of the source, $f_{\rm esc}$ is the escape fraction of photons from the source, $f_{\rm star}$ is the fraction of gas converted to stars, and ϵ_ν is the composite stellar emissivity per unit mass.

It is instructive to examine the evolution of the ionization front in units of the time–dependent Strömgren radius of a steady source at the initial luminosity,

$$r_{\rm s} = \left[\frac{3}{4\pi\alpha_{\rm B} n_{\rm H}^2(t)} \frac{dN_\gamma}{dt}(t=0)\right]^{1/3}. \tag{5.8}$$

Figure 5.7 illustrates the evolution of the ionization front in units of the Strömgren radius for sources that turn on at different redshifts between 10–100, and have the ionizing photon production rate shown in Figure 5.4. Two interesting effects are apparent:

1. The ionization front reverses its direction of propagation when the flux from the source starts to decline (cf. Fig. 5.4). The decrease in the volume of the HII region can be substantial at high redshifts, when the recombination time in the IGM is short.

2. At redshifts $z \lesssim 10^2$, the ionization front does not have sufficient time to expand to its full Strömgren radius, as pointed out by Shapiro and Giroux (1987). This effect is less important at high redshifts, because the recombination time in the IGM is then shorter and the Strömgren radius is therefore closer to the source.

Note that our calculation assumes that the ionization front propagates out from a virialized nonlinear object into a uniform IGM. Numerical simulations (such as described in Gnedin & Ostriker 1996) are necessary in order to explicitly incorporate the mildly nonlinear IGM clumpiness in this calculation.

5.2.4 Carbon Production and Star Formation Efficiency

Because we presently lack a standard theory of star formation, our models (as well as other studies, such as Fukugita & Kawasaki 1994, and Madau & Shull 1996) simply assume that a constant fraction $f_{\rm star}$ of the gas is transformed into stars. Our approach is to calibrate $f_{\rm star}$ based on the inferred carbon abundance in the IGM from Keck observations of the Lyα forest (Cowie et al. 1995; Tytler et al. 1995; Songaila & Cowie 1996, Cowie 1996). The low column density absorbers in the Lyα forest extend over large transverse scales of $\gtrsim 0.1$–1 Mpc (Bechtold et al. 1994; Dinshaw et al. 1994, 1995; Fang et al. 1995; Smette et al. 1995) and can be interpreted as mildly overdense regions in a photoionized IGM (Hernquist et al. 1996, Miralda-Escudé et al. 1995). We assume that the apparently universal metallicity of these systems resulted from metal enrichment by many low–mass ($\sim 10^{8-9} M_\odot$) star clusters inside of them. In our standard model we adopt the most conservative approach which assumes that all the metals produced by the stars are mixed efficiently with the IGM;

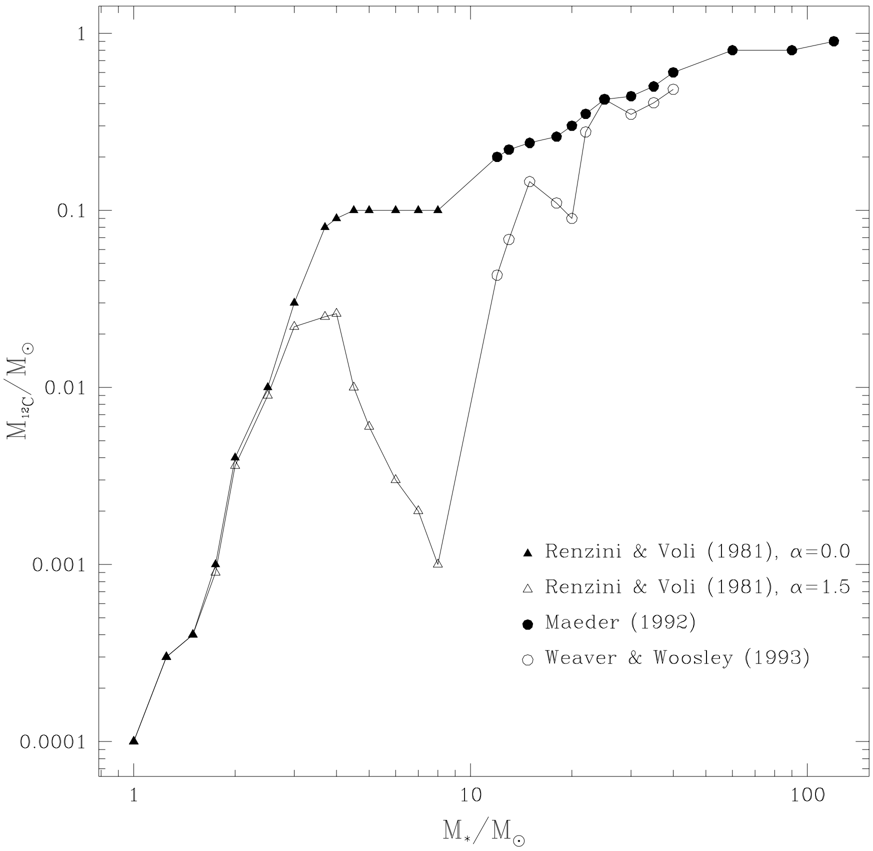


Figure 5.8: *Carbon yields* ($M_{^{12}C}$ *in solar masses) from stars of different masses. The large allowed range in the yields between 3–8* $M_\odot$ *in Renzini & Voli (1981) is due to the uncertainty in the extent of hot bottom burning;* $\alpha = 1.5$ *and 0 correspond to efficient, or no hot bottom burning, respectively.*

a lower mixing efficiency would imply a higher value for $f_{\rm star}$ and an even earlier reionization epoch for the same value of IGM metallicity. The required mixing of the metal rich gas with the IGM could have resulted from supernova driven winds, since the velocity dispersion of the systems which are responsible for reionization in our models is only ~ 10 km s^{-1}. However, even if some metallicity gradients remained because of incomplete mixing, the line-of-sight integral inherent to observations of the Lyα forest systems tends to average over these variations and yield a roughly universal metallicity value throughout the universe.

We assume that the same population of stars that produced the heavy elements detected in the Lyα absorbers is responsible for the reionization of the universe. In order to fix $f_{\rm star}$, we need to specify the amount of heavy elements produced by a population of stars with the Scalo IMF. Since the Keck observations detected specifically CIV absorption lines, we focus our attention on stellar yields of ^{12}C. The carbon yields of high mass stars are well established in the literature. Weaver & Woosley (1993) have calculated the yields for stars with masses between 11–40$M_\odot$, and their numbers are in good agreement with the yields reported by Maeder (1992) for stars with 9–120$M_\odot$. For low and intermediate mass stars with 1–8$M_\odot$, the yields were tabulated by Renzini & Voli (1981), and more recently by van den Hoek & Groenwegen (1996). However, since these stars do not explode as supernovae, there

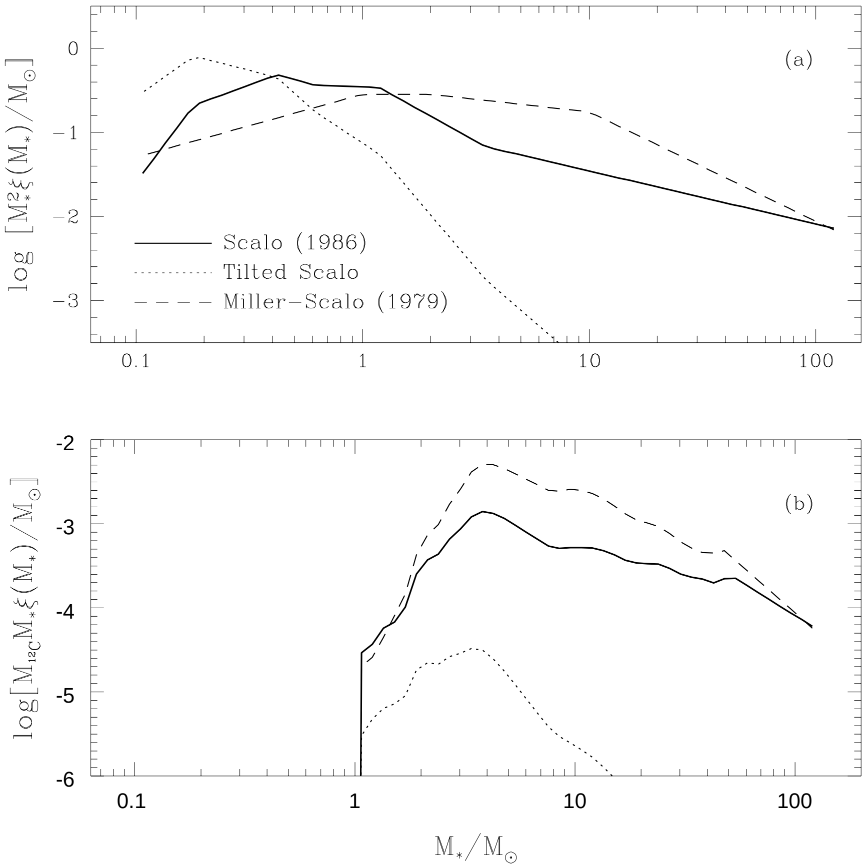


Figure 5.9: *Upper panel: three different initial mass functions (IMFs). The Scalo (1986) and Miller-Scalo (1979) IMFs are two different parameterizations of the locally observed stellar population, and the tilted Scalo IMF is obtained by adding a constant β=1.69 to the power law index in each segment of the Scalo IMF, while keeping the value of the IMF fixed at $4M_\odot$. All three IMFs are normalized to an integral of $1M_\odot$. Lower panel: carbon yields from stars of different masses (*M$_{^{12}\mathrm{C}}$*, in solar masses), folded in with the three different IMFs. In all cases, most of the carbon is produced by $\sim$ 3–6$M_\odot$ stars.*

is a significant uncertainty about the amount of carbon that actually gets ejected from these stars to the surrounding medium (e.g. through thermal pulsations or winds). This uncertainty is particularly large for stars which are more massive than $\sim 3M_\odot$, as these stars are known to undergo some amount of "hot bottom burning", i.e. burning of carbon into nitrogen at the base of their convective envelopes. The actual amount of hot bottom burning depends sensitively on the convective mixing length, and is highly uncertain.

Figure 5.8 shows the carbon yields as a function of stellar mass. The large range of yields allowed for 3–8$M_\odot$ stars is due to the unknown efficiency of hot bottom burning. Additional uncertainty in the total carbon yield results from the unknown stellar IMF; Figure 5.9 shows the *maximum* allowed carbon yields folded in with a Scalo (1986), Miller–Scalo (1979), and tilted Scalo IMFs. In our standard model we adopt the first choice. The sensitivity of our results to variations in these uncertain parameters will be examined in § 4.

Figures 5.8 and 5.9 show that most of the carbon is produced by 3–6$M_\odot$ stars, and therefore the total amount of carbon output by all stars is subject to the uncertainties about hot bottom burning. In order to get an average carbon enrichment equivalent to 1%$Z_\odot$, as measured in the Lyα forest, it is necessary that a nominal 2.8% of all the baryons be converted into stars with a Scalo IMF (or a fraction 1.2% for a Miller–Scalo IMF). The corresponding required value of the star–formation efficiency for the Scalo IMF is $f_{\rm star} = 13\%$. Here, a factor of ~ 2.3 increase is from the average time required to produce carbon inside the stars; i.e. only a fraction 1/2.3 of the total stellar carbon is produced by $z = 3$; another factor of 2 increase is from the value of the collapsed fraction, $F_{\rm coll} = 50\%$ at $z = 3$. Note that $f_{\rm star}$ must be increased further if any significant amount of hot bottom burning occurs in stars with 3–8$M_\odot$.

Finally, it is interesting to note that most of the carbon enrichment is due to stars with 3–6$M_\odot$, while most of the ionizing photons are produced by stars with 10–12$M_\odot$. Because of this fact, the carbon enrichment is delayed considerably relative to the photon production; 90% of ionizing photons are emitted already within $\sim 5 \times 10^7$ years following the starburst, while it takes $\sim 5 \times 10^8$ years to eject 90% of the total carbon output.

5.3 Consequences of the Model

Next, we use the model described in the previous section to calculate the redshift evolution of various integral quantities, i.e. the fraction of baryons contained in collapsed objects, the average metallicity of the IGM, the volume filling factor of ionized regions, the microlensing probability, and the optical depth to electron scattering which translates to a damping factor for the cosmic microwave background (CMB) anisotropies. Below we discuss each of these quantities separately.

The Collapsed Fraction

The fraction of baryons contained in collapsed objects is given in equation (5.1). In our model, a fixed universal fraction $f_{\rm star}$ of this material is converted into stars. However, as argued in §2, fragmentation of the collapsed clouds into stars is initially inhibited by the photo–dissociation of H_2 (cf. HRL97). Shortly after the very first ionizing sources turn on, H_2 cooling is suppressed even in dense environments. Star formation resumes when clouds with a virial temperature $\gtrsim 10^4$K, or a total mass $\gtrsim 10^8 M_\odot[(1+z)/10]^{-3/2}$ collapse, inside which atomic line cooling becomes efficient and triggers fragmentation. We incorporate this pause in the star-formation history into our model by setting $f_{\rm star} = 0$ in equation (5.1) for clouds with masses below $10^8 M_\odot[(1+z)/10]^{-3/2}$. Note, however, that star formation is suppressed by a factor of $\sim [1 - F_{\rm HII}(z)]$ even in the absence of H_2-feedback; the collapse of clouds in a hot ($\sim 10^4$ K) photoionized gas is delayed until dark-matter potential wells with sufficient depth form (Couchman & Rees 1986). We therefore multiply $f_{\rm star}$ by $[1 - F_{\rm HII}(z)]$ in one of the variants on our standard model where H_2-feedback is ignored.

Metallicity

We obtain the average metallicity of the IGM from the carbon yields as described in § 2.4. Our standard model assumes that each star deposits all of its carbon yield in the background gas at the end of its life on the main sequence, and that the metal enriched gas is fully mixed with the IGM[2]. We define $f_{\rm carb}(\Delta t)$ to be the fraction of the total carbon yield which gets deposited after a time Δt following the starburst. As a result of this deposition, the IGM obtains a carbon metallicity

$$Z(z) = 0.01 Z_\odot \times \left(\frac{f_{\rm star}}{0.028}\right) \int_\infty^z dz' \frac{dF_{\rm coll}}{dz}(z') f_{\rm carb}\left[t(z) - t(z')\right]. \tag{5.9}$$

The value $f_{\rm star} = 0.028$ would result in the required metallicity $0.01 Z_\odot$ if the collapsed fraction $F_{\rm coll}$ was unity and there was no time–delay in the carbon prodcution (assuming the maximum possible carbon yields with a Scalo IMF, cf. § 2.4). $Z(z)$ is proportional to $F_{\rm coll}(z)$ only at low redshifts when the finite time needed to deposit the carbon yields is much shorter than the Hubble time.

HII Filling Factor

Our model can be used to determine the fraction of the volume of the universe in which hydrogen is ionized as a function of redshift. We define $r_{\rm i}(z_{\rm on}, z, M)$ to be the time dependent radius of the spherical ionization front around a source of mass M which turns on at a redshift $z_{\rm on}$. Equation (5.6) can then be used to determine $r_{\rm i}(z_{\rm on}, z, M)$ for a range of z and M values. Since the escape fraction of ionizing photons $f_{\rm esc}$ is, to leading order, independent of the mass of the system, the volume ionized by a given system scales linearly with its mass, namely $V_{\rm HII}(z_{\rm on}, z, M) = (4\pi r_{\rm i}^3/3) = M \times \tilde{V}_{\rm HII}(z_{\rm on}, z)$, where $\tilde{V}_{\rm HII}(z_{\rm on}, z)$ is the ionized volume per unit mass. We can therefore evaluate the HII filling factor $F_{\rm HII}(z)$ from the collapsed gas fraction

$$F_{\rm HII}(z) = \rho_{\rm b}(z) \int_\infty^z dz' \frac{dF_{\rm coll}}{dz}(z') \tilde{V}_{\rm HII}(z', z), \tag{5.10}$$

where $\rho_{\rm b} = \Omega_{\rm b} \rho_{\rm crit}$ is the average baryonic density and $\rho_{\rm crit}$ is the critical density of the universe. Note that this result depends only the total fraction of baryons in nonlinear objects and not on the merger history of these objects.

Microlensing Optical Depth

The probability for having a microlensing event along the line-of-sight to a point source located at a redshift $z_{\rm S}$ in an $\Omega_0 = 1$ universe, is given by,

$$\tau_{\rm lens}(z_{\rm S}) = \frac{\Omega_{\rm b} \rho_{\rm crit} c}{H_0} \int_o^{z_{\rm S}} dz_{\rm L} \sqrt{1 + z_{\rm L}} f_{\rm star} F_{\rm coll}(z_{\rm L}) \sigma_{\rm lens}(z_{\rm L}), \tag{5.11}$$

where c is the speed of light, and the cross–section per unit mass associated with the Einstein ring of the lenses has the form (Gould 1995; Turner, Ostriker & Gott 1984):

$$\sigma_{\rm lens}(z_{\rm L}) = \frac{4\pi G}{c^2} \frac{x_{\rm L}(\lambda_{\rm S} - \lambda_{\rm L})\lambda_{\rm L}}{\lambda_{\rm S}}. \tag{5.12}$$

Here the subscripts L and S refer to the lens and the source, $x \equiv (1 + z)$, and

$$\lambda \equiv \frac{2}{5} \frac{c}{H_0} \left(1 - x^{-5/2}\right). \tag{5.13}$$

We implicitly assume that the intergalactic star clusters are intrinsically optically thin to microlensing.

[2]A more refined modeling might treat the dependence of the "mixing efficiency" with the IGM on the sizes of the aggregates in which stars form (Rees 1996).

Optical Depth to Electron Scattering

The optical depth to electron scattering out to a redshift z in an $\Omega_0 = 1$ universe, is given by,

$$\tau_{\rm es}(z) = 0.053\Omega_{\rm b}h\int_0^z dz'\sqrt{1+z'}\left[1 - f_{\rm star}F_{\rm coll}(z')\right]F_{\rm HII}(z'), \qquad (5.14)$$

where we used the fact that the average ionization fraction is equal to the HII filling factor. We have implicitly assumed that the fraction of singly ionized helium is $F_{\rm HII}$, and that no helium is ionized twice due to the sharp drop in our template spectrum at the HeII edge (cf. Fig. 5.3).

CMB Anisotropy Damping

The CMB anisotropies are damped on scales smaller than the apparent angular size of the horizon at the reionization epoch ($\sim 10°$), due to scattering off free electrons. Hu & White (1996) provide a fitting formula for the damping factor of the CMB power–spectrum, R_l^2, as a function of the index l in the spherical harmonic decomposition of the microwave sky [the angular scale corresponding to a given l–mode is $\sim 1° \times (l/200)^{-1}$]. The damping factor is uniquely related to the scattering optical depth as a function of redshift, $\tau_{\rm es}(z)$, which is given by equation (5.14) (see also Kamionkowski et al. 1994).

5.4 Results and Discussion

Figures 5.10–5.15 show results from our reionization models. We define our "standard model" by the set of parameters summarized in Table 5.1, and analyze the sensitivity of the results to separate changes in the values of each parameter. Our choice for the cosmological parameters follows standard CDM with a scale–invariant power spectrum, and the likely nucleosynthesis value for the baryon density (Copi, Schramm & Turner 1995), namely n=1, h=0.5, Ω = 1, $\Omega_{\rm b}$=0.05. The COBE normalization of this power spectrum implies $\sigma_{8h^{-1}}$=1.22, which predicts too much power on small scales and is ruled out by observations (Bunn & White 1996). We therefore use as our standard value $\sigma_{8h^{-1}}$=0.67 (an upper limit based on cluster abundance, see White, Efstathiou, & Frenk 1993; Bond & Myers 1996; Viana & Liddle 1996). It is important

Table 5.1: *The assumed parameter values in our standard model and its variants.*

Parameter	Standard Model	Range Considered
$\sigma_{8h^{-1}}$	0.67	0.67–1.0
n	1.0	0.8–1.0
$\Omega_{\rm b}$	0.05	0.01–0.1
$f_{\rm star}$	13%	1%–40%
$f_{\rm esc}$	$f_{\rm esc}(z)$	3%–100%
IMF tilt (β)	0	0–1.69
H_2 feedback	yes	yes/no

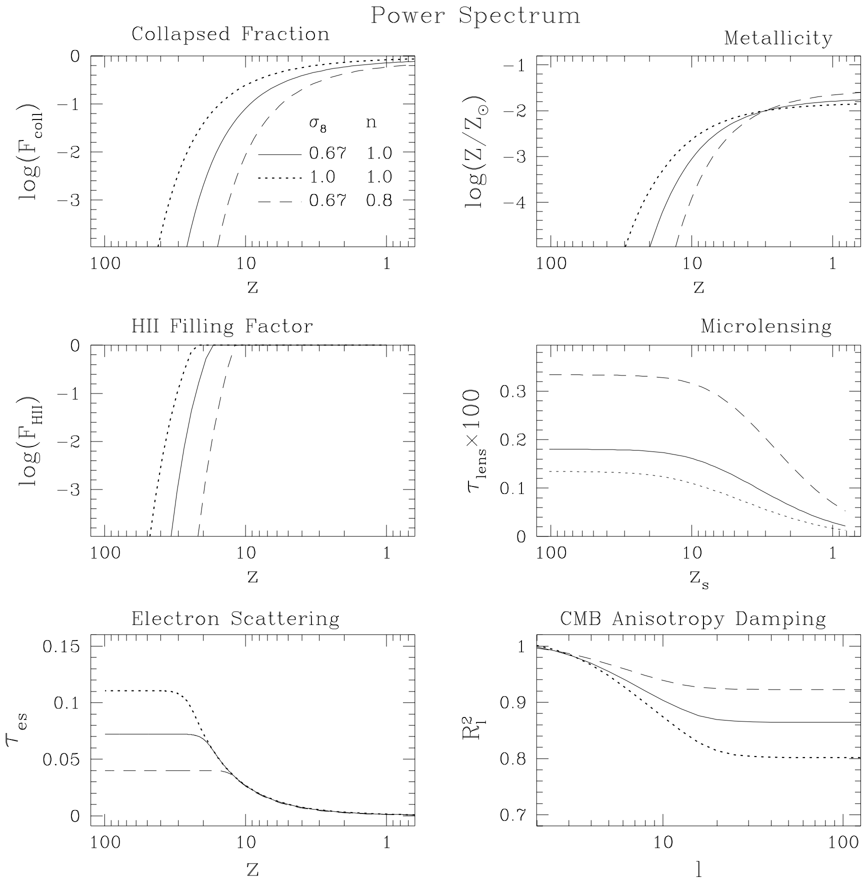


Figure 5.10: *The different panels show (i) the collapsed fraction of baryons; (ii) the average metallicity of the IGM; (iii) the volume filling factor of ionized regions; (iv) the optical depth to microlensing; (v) the optical depth to electron scattering; and (vi) the net damping factor for the power–spectrum decomposition of CMB anisotropies as a function of the spherical harmonic index l. In this figure we examine the sensitivity of the results to changes in the normalization $\sigma_{8h^{-1}}$ or the primordial index n of the CDM power–spectrum.*

to note that our results depend primarily on the behavior of the power spectrum on small scales and have little sensitivity to its characteristics on large scales, where direct constraints from galaxy surveys and microwave anisotropy experiments apply. In our standard model, we adopt a Scalo stellar IMF with the maximum carbon output (i.e. the minimum amount of ionizing photons for a given metallicity), and assume the existence of a negative feedback on star–formation due to photodissociation of H_2. The standard star formation efficiency $f_{\rm star}$ is chosen to be 13%, so as to yield a metallicity of $0.01 Z_\odot$ at a redshift $z = 3$, where the collapsed fraction is $F_{\rm coll} \approx 50\%$ (cf. §2.4). The standard escape fraction $f_{\rm esc}$ is explicitly calculated at each redshift as described in § 2.2 and shown in Figure 5.6.

The results from our standard model are described by the solid lines in Figures 5.10–5.15. In this model, reionization occurs at $z = 18$, with a corresponding scattering optical depth of $\tau_{\rm es} = 0.07$. This leads to a $\sim 14\%$ reduction in the power of CMB anisotropies on angular scales $\lesssim 10°$. The microlensing optical depth is small, $\tau_{\rm lens} \lesssim 2 \times 10^{-3}$, even at high redshifts. Figure 5.10 also shows results for a

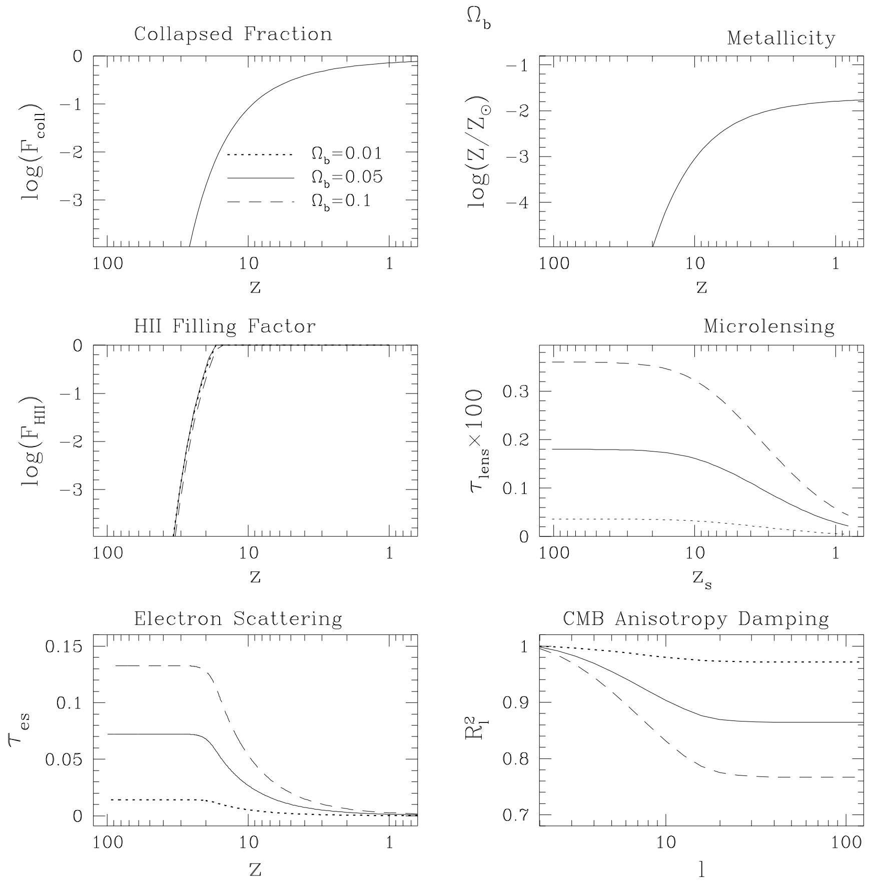


Figure 5.11: *Same as Figure 5.10, but showing the sensitivity of the results to different values of the baryon density parameter,* Ω_b.

higher normalization ($\sigma_{8h^{-1}}$=1.0) and a tilt (n=0.8) of the CDM power spectrum. With these variations, the reionization redshift changes to z=22 and z=13, and the scattering optical depth changes to $\tau_{es} = 0.11$ and $\tau_{es} = 0.04$, respectively. The low level of these changes demonstrates that our results are robust for popular variants of the CDM power spectrum.

Next, we consider the dependence of the results on the baryon density. Figure 5.11 shows the cases of Ω_b=0.01 and 0.1. The microlensing optical depth is simply proportional to Ω_b, while the electron scattering optical depth for these cases is $\tau_{es} = 0.02$ and $\tau_{es} = 0.13$, respectively. The deviation from the exact proportionality $\tau_{es} \propto \Omega_b$ is due to the increased recombination rate in a high baryon density universe, which slightly decreases the reionization redshift.

Figure 5.12 illustrates the significance of changing the star formation efficiency f_{star}. Although f_{star} is directly constrained by the observed IGM metallicity, there are two reasons to consider variations in this parameter. First, there is approximately an order of magnitude uncertainty in the observed value of the CIV/H ratio; this makes the metallicity and therefore f_{star} uncertain by the same factor. Second, if hot bottom burning is efficient, it could eliminate the carbon yields from 3–8$M_\odot$ stars, and increase f_{star} by up to a factor of 4. We therefore examine the cases f_{star}=1% and f_{star}=40%.

For f_{star}=40% the reionization redshift is increased to z=24, but (i) the reion-

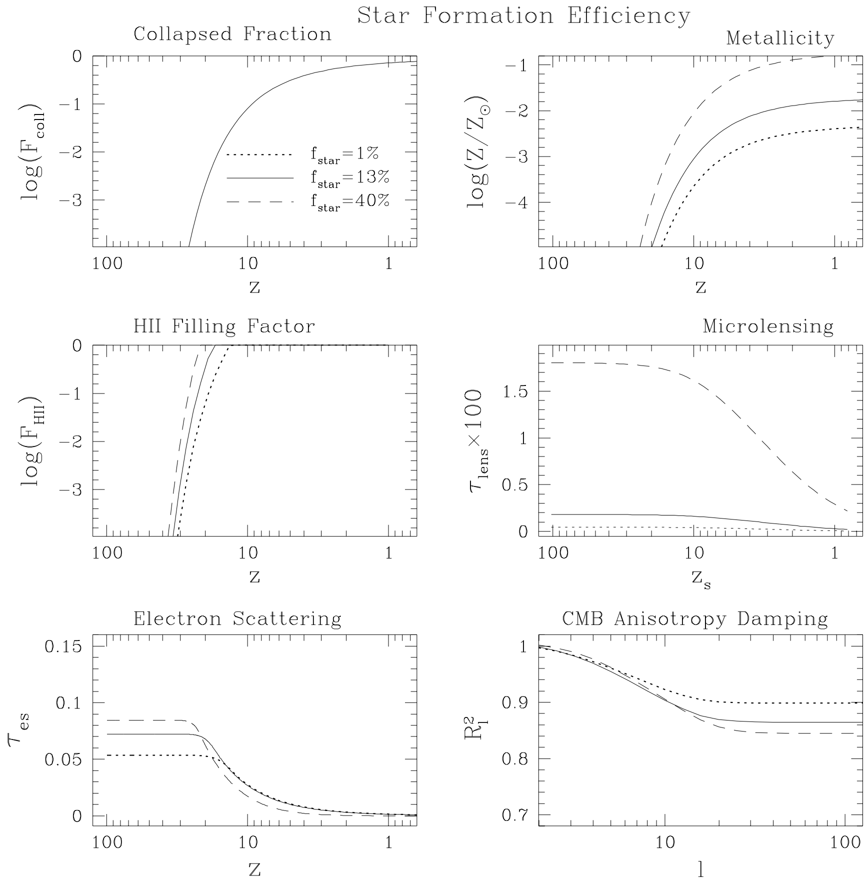


Figure 5.12: *Same as Figure 5.10, but showing the sensitivity of the results to a change in the star formation efficiency* $f_{\rm star}$*. If the carbon output from the stars is kept fixed, then a change in* $f_{\rm star}$ *translates to a proportional change in the metallicity, as shown in the upper right panel. However, the carbon yield from stars is uncertain, and thus for the other panels–an increase of* $f_{\rm star}$ *by some factor may also be viewed as a decrease in the carbon yield by the same factor, while keeping the IGM metallicity fixed.*

ization in this case is more sudden; and (ii) a substantial fraction of the electrons are locked up in stars and do not contribute to $\tau_{\rm es}$ (the gas fraction is reduced by the factor $[1 - f_{\rm star}F_{\rm coll}]$, cf. eq. [5.14]). These effects counteract the increase in the reionization redshift, and yield an optical depth, $\tau_{\rm es} = 0.09$, which is only 15% higher than with $f_{\rm star}$=13%. On the other hand, $f_{\rm star}$=1% gives $\tau_{\rm es}$=0.05, again not too different from $\tau_{\rm es}$=0.07. Hence, Figure 5.12 demonstrates that the scattering optical depth is fairly insensitive to large variations in the carbon output.

Figure 5.13 demonstrates the effect of varying the IMF. Although the stellar IMF at high redshift is unknown, it is often claimed that due to the absence of metals and the lack of initial substructure, it favored massive stars (e.g. Carr, Bond, & Arnett 1984). For this reason, we show in Figure 5.13 the results for the Miller–Scalo IMF (dotted lines), which is biased towards massive stars relative to the standard Scalo IMF (solid lines). Note that the bias towards massive stars increases both the carbon yield and the UV photon production per solar mass. Therefore, a smaller number of stars are necessary to provide the observed metallicity ($f_{\rm star} = 5\%$), but

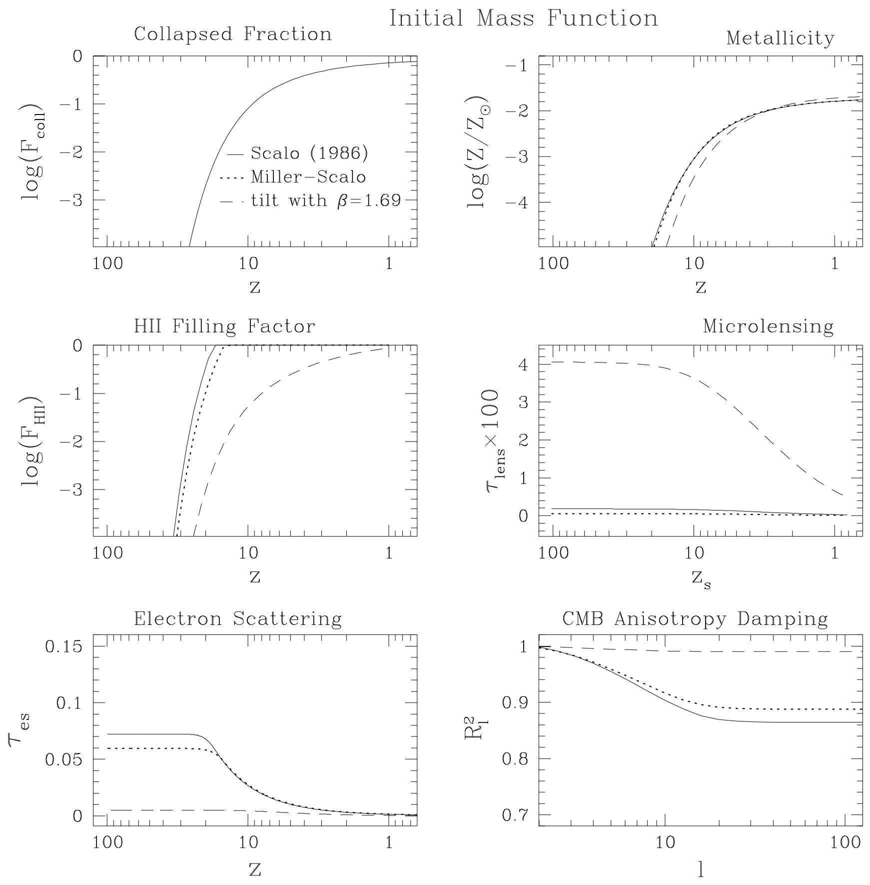


Figure 5.13: *Same as Figure 5.10, but showing the effect of tilting the Scalo IMF, i.e. making it steeper by adding a constant β=1.69 to the power law index in each segment. Results are also shown for a Miller–Scalo IMF which shows the opposite bias in that it contains more high mass stars than the Scalo IMF.*

these stars produce more ionizing photons. As the figure shows, the two effects almost cancel each other, and reionization with the Miller–Scalo IMF occurs at $z = 14$, close to the redshift with the Scalo IMF.

On the other hand, the possibility still exists that the high–redshift IMF is biased in the opposite sense, i.e. towards low–mass stars. Indeed, in order to identify the early stars with MACHOs which contribute significantly to the mass of galaxy halos, a substantial fraction of the baryons must be converted into stars, and most of these stars must have low masses, $\lesssim 0.5\mathrm{M}_\odot$ (Flynn et al. 1996; Alcock et al. 1996). We therefore assume $f_{\rm star} = 90\%$. The tilt necessary to increase $f_{\rm star}$ from 13% to 90% while keeping the metallicity fixed, must decrease the total carbon yield per unit mass by a factor of 60. This is achieved by adding a constant $\beta = 1.69$ to the power–law index in all segments of the Scalo IMF. This tilt decreases the average stellar mass from $0.5\mathrm{M}_\odot$ to $0.25\mathrm{M}_\odot$ (cf. Fig. 5.9). The dashed lines in figure 5.13 shows that with this tilt, stellar reionization is delayed substantially down to $z \sim 1$ (at which point quasars, which are ignored in this work, dominate the ionizing background). This delay is caused by the lack of ionizing photons from massive stars and is therefore a robust conclusion; combining a tilt with $f_{\rm esc} = 100\%$ only increases the reionization redshift up to $z \sim 2$.

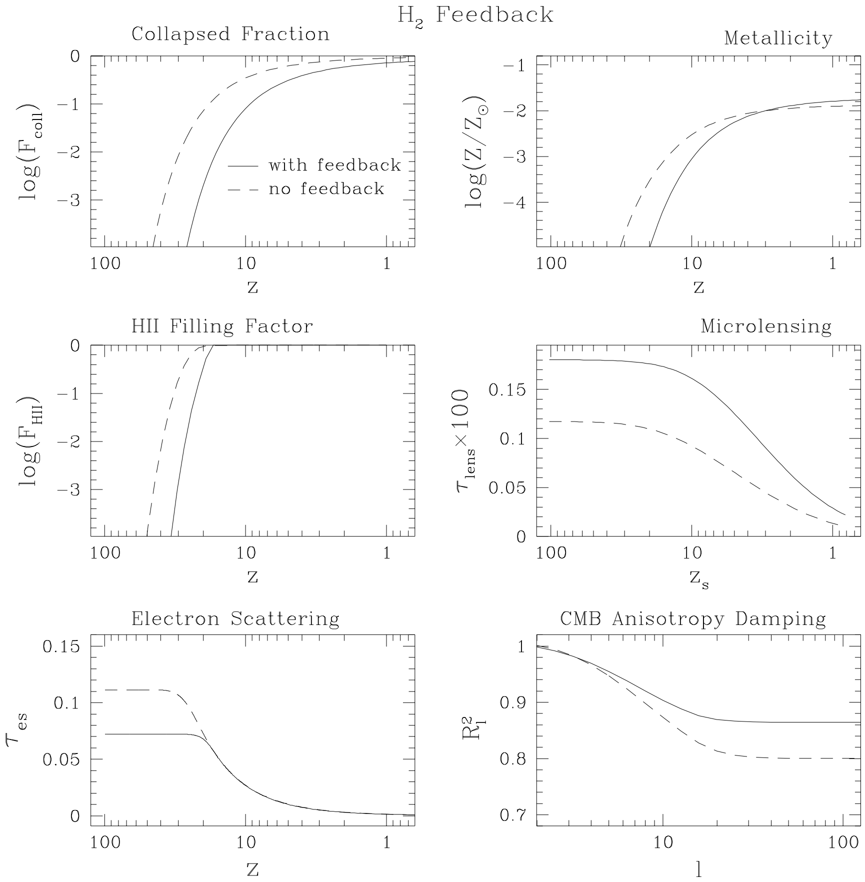


Figure 5.14: *Same as Figure 5.10, but showing the significance of the negative feedback on star formation due to the photodissociation of* H_2.

According to Figure 5.13, a tilted IMF increases the optical depth to microlensing substantially up to values as high as 3% at z=5 or 4% at $z \gtrsim 30$. Such microlensing probabilities could leave detectable signatures in large samples of high redshift quasars. Microlensing could either be detected through measurements of the variability of quasars or the systematic changes in the equivalent width of their emission lines. The variability signature requires an impractically long observing period (of order tens of years for solar mass lenses), and is difficult to separate from an intrinsic variability signal. The equivalent–width change results from the fact that a stellar lens could magnify the quasar continuum source, which is typically smaller than its Einstein radius $\sim 10^{16}(M_\star/0.1M_\odot)^{1/2}$ cm, but would not magnify the broad line region which is much more extended (see Dalcanton et al. 1994, and references therein).

Figure 5.14 shows the results with and without the H_2-photodissociation feedback on star-formation. Even though the collapsed baryonic fraction is much higher without this feedback, in both cases reionization occurs at about the same redshift, $z \approx 20$. The similarity in the reionization redshifts results from the fact that when the HII filling factor approaches unity, star formation is suppressed both with and without the H_2 feedback (cf. § 3). The total optical depth to electron scattering without the H_2–feedback is increased to $\tau_{es} = 0.11$; most of the difference from the case with the H_2–feedback is introduced when the universe is only partially ionized.

Figure 5.15 illustrates the sensitivity of the results to the escape fraction. We

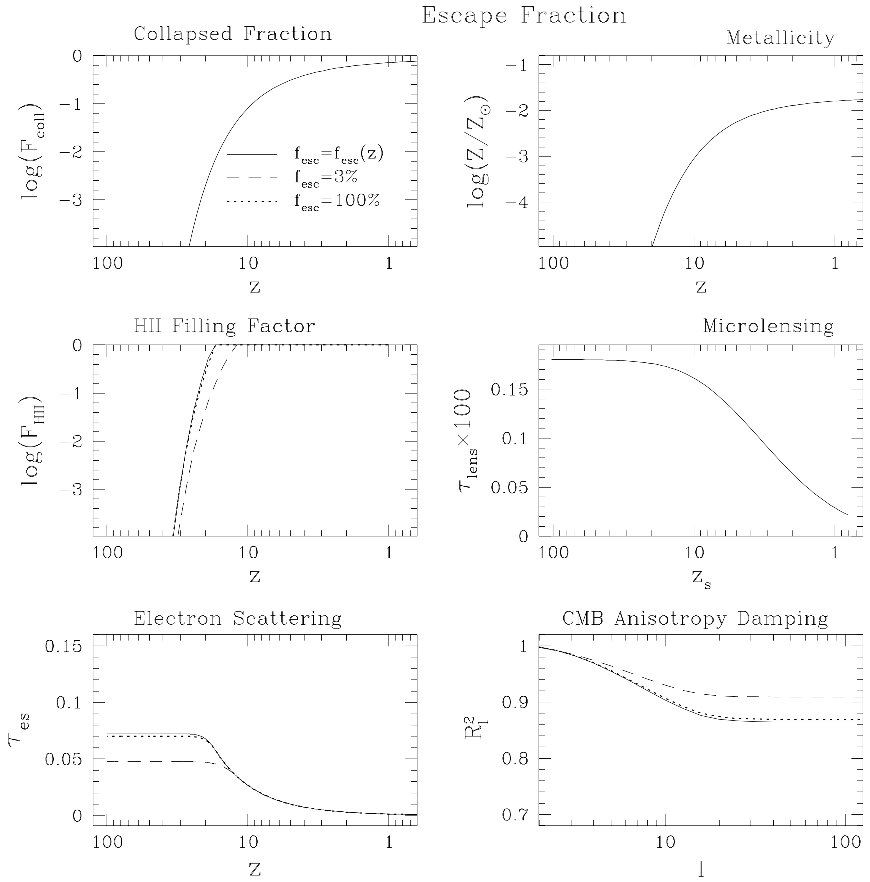


Figure 5.15: *Same as Figure 5.10, but showing the sensitivity of the results to changes in the escape fraction of ionizing photons, $f_{\rm esc}$.*

show results for either the low value of $f_{\rm esc}$=3% observed in nearby starburst galaxies (Leitherer et al. 1996), or the hypothetical high value of 100%. The motivation for considering $f_{\rm esc}$=100% is the theoretical possibility that the obscuring gas could be expelled from the star forming clouds by photoionization heating or supernova–driven winds. Such an expulsion of the gas would then quench further star-formation in these systems (Lin & Murray 1992); this mechanism could explain why the star formation efficiency is limited to 13%. In this work, however, we normalize the star formation efficiency based on the metallicity of the Lyα forest, without attempting to explain its origin. In the case of a low escape fraction, the resulting optical depth for electron scattering is decreased to $\tau_{\rm es}$=0.05, with reionization still occurring at z=11. In the case of a high escape fraction, the optical depth and reionization redshifts are increased by a negligible fraction.

In addition to the above indirect signatures of reionization, it is interesting to examine the possibility of a direct imaging of the reionization sources. Although any such imaging is beyond the sensitivities of current instruments, the Next Generation Space Telescope (NGST) will achieve a suitable sensitivity for that purpose, $\sim$ 1 nJy in the wavelength range 1–3.5μm (Mather & Stockman 1996; see also http://ngst.gsfc.nasa.gov). In Figure 5.16 we show the predicted number of objects that NGST would detect as a function of observed flux, assuming a sudden reionization at the redshift $z_{\rm ion}=18$. The number counts are normalized to the field of view of the Space Infrared Telescope Facility (SIRTF) $5' \times 5'$, which is scheduled for launch long before NGST (the planned field-of-view of NGST is similar, $4' \times 4'$). The abun-

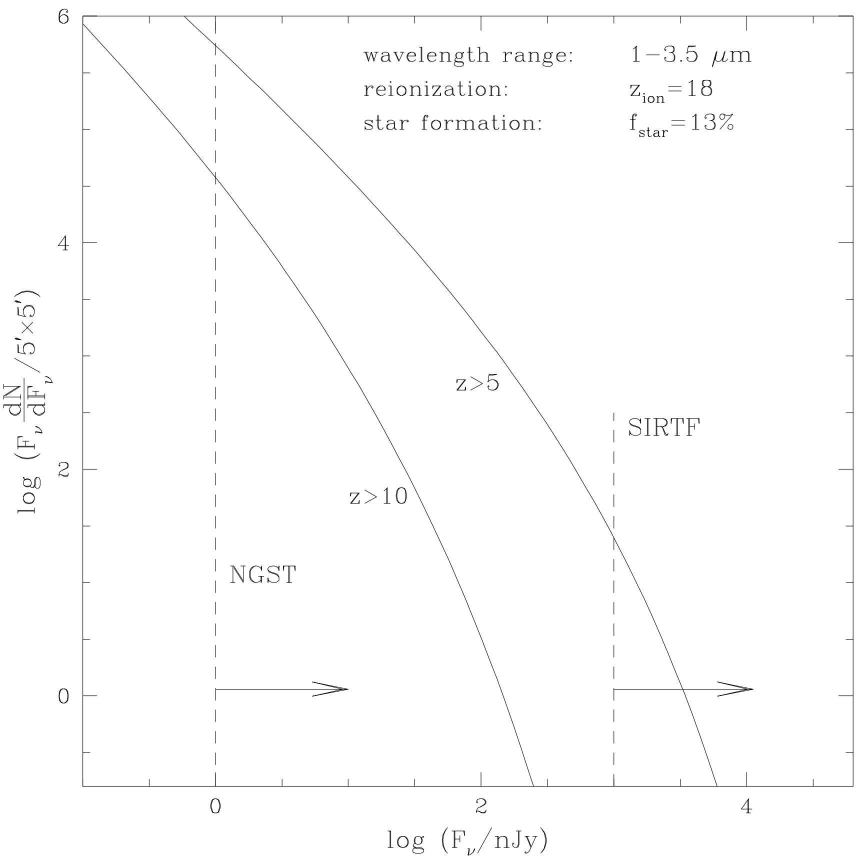


Figure 5.16: *The number per logarithmic flux interval of high–redshift objects which could be probed by future space telescopes, in the wavelength range of 1–3.5μm. In this calculation, we assume sudden reionization at z=25. The vertical dashed lines show the expected sensitivities of the Space Infrared Telescope Facility (SIRTF) and the Next Generation Space Telescope (NGST).*

dance and emissivity of each source was calculated according to our standard model as described in § 2.1-2.2. We have averaged the flux from each source over the range of 1–3.5μm and truncated the spectra of all objects at emission frequencies above Lyα before reionization occurs, since the optical depth of HI at these frequencies is exceedingly high prior to reionization.

Figure 5.16 shows the number per logarithmic flux interval of all objects at $z > 5$ (top curve), and all objects at $z > 10$ (bottom curve). The number of detectable sources is high. With its expected sensitivity, NGST will be able to probe about 10^4 objects at $z > 5$. The average separation between these compact sources would be of order 2.4″, well above the angular resolution of 0.06″, planned for NGST.

Another detectable signature of a reionized IGM results from its Bremsstrahlung emission. In an $\Omega = 1$ universe, the surface brightness observed today is given by (Loeb 1996)

$$J_{\rm ff}(\nu) = \frac{c}{H_0}\int_\infty^0 \frac{j_{\rm ff}[(1+z)\nu]dz}{(1+z)^{11/2}}, \tag{5.15}$$

where the free–free emissivity $j_{\rm ff}$ at very low frequencies is independent of frequency

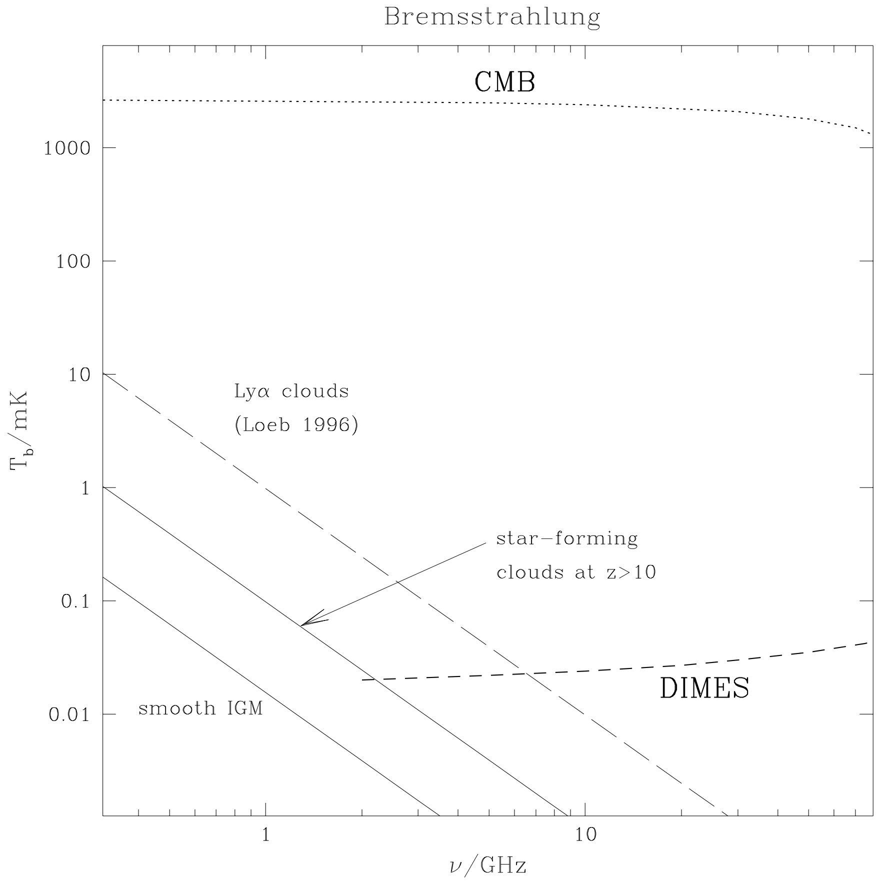


Figure 5.17: *Observed brightness temperature from Bremsstrahlung emission by star–forming clouds at* $z > 10$ *and ionized regions of the smooth IGM in our standard model. The emission from the star–forming clouds exists only if the clouds retain their gas over the lifetime of their stellar population. For comparison, we show the Bremsstrahlung emission from the Ly*α *forest between* $0 < z < 5$ *for a UV background of* 10^{-21} erg cm^{-2}s^{-1}Hz^{-1}sr^{-1} *at the Lyman–limit (Loeb 1996). We also show the CMB spectrum, and the proposed sensitivity of the DIMES experiment (Kogut 1996; see also http://ceylon.gsfc.nasa.gov/DIMES).*

(we ignore the logarithmic frequency–dependence of the Gaunt factor), and is given by (Rybicki & Lightman 1979)

$$j_{\rm ff}(\nu) = 3.1 \times 10^{-40} \frac{\langle n_{\rm e}^2 \rangle}{(T/10^4{\rm K})^{0.5}} \quad {\rm erg\ cm^{-3}\ s^{-1}\ Hz^{-1}\ sr^{-1}}. \tag{5.16}$$

The value of $\langle n_{\rm e}^2 \rangle$ at each redshift is a sum of the contributions from the ionized regions of the smooth IGM and the star–forming clouds themselves,

$$\langle n_{\rm e}^2 \rangle = F_{\rm HII}(1 - F_{\rm coll}) \langle n_{\rm e}^2 \rangle_{\rm IGM} + \int_\infty^z dz' \frac{dF_{\rm coll}}{dz}(z') \langle n_{\rm e}^2 \rangle_\star [t(z) - t(z')]. \tag{5.17}$$

In our standard model, the average value of $n_{\rm e}^2$ within an ionized region of the smooth IGM is simply given by $\langle n_{\rm e}^2 \rangle_{\rm IGM} = [1.15 \times 10^{-7}(1+z)^3]^2$ cm^{-6}. The average of $n_{\rm e}^2$ within each star forming cloud can be obtained by noticing that the recombination rate in each cloud is simply $(1 - f_{\rm esc})$ times the production rate of ionizing photons, i.e.

$$\langle n_{\rm e}^2 \rangle_\star [\Delta t] \equiv \frac{1}{\frac{4}{3}\pi R^3} \int_0^R 4\pi r^2 dr n_{\rm e}^2(r) = \frac{1}{\alpha_{\rm B}} \frac{1}{\frac{4}{3}\pi R^3} \left(\frac{1 - f_{\rm esc}}{f_{\rm esc}} \right) \frac{dN_\gamma}{dt} [\Delta t]. \qquad (5.18)$$

This result for the clumpiness factor is independent of the cloud mass and is only a function of the time lag since the starburst occurred, Δt. The escape fraction $f_{\rm esc}$ in this expression was derived in § 2.2 (cf. Fig. 5.6), and the time–dependent production rate of ionizing photons dN_γ/dt was calculated in § 2.3 (cf. eq. 5.7 and Fig. 5.4).

The resulting brightness temperature, $T_{\rm b} \equiv (c^2 J_{\rm ff}/2k\nu^2)$, of the microwave sky due to Bremsstrahlung emission by star–forming clouds at $z > 10$ and the smooth IGM, is shown in Figure 5.17 for the frequency range $\nu = 1$–100 GHz. The existence of Bremsstrahlung emission from star forming regions relies on their ability to retain their gas during the lifetime of their stellar population, and is therefore uncertain. Reprocessing of starlight by dust (Wright 1981, Wright et al. 1994) might add to this emission component and will be considered elsewhere (Loeb & Haiman 1997). For comparison, the long–dashed curve shows the unavoidable Bremsstrahlung signal from the Lyα forest clouds (with HI column densities $\lesssim 10^{17}$ cm^{-2}) at $0 \lesssim z \lesssim 5$, for a UV background flux of 10^{-21} erg cm^{-2}s^{-1}Hz^{-1}sr^{-1} at the Lyman–limit (Loeb 1996). The calculation of the Lyα forest contribution avoids any uncertainties about the clumpiness of the IGM, because at photoionization equilibrium the line-of-sight integral of $n_{\rm e}^2$ is proportional to the *observed* column density of neutral hydrogen. The above signals are compared to the expected sensitivity of the proposed Diffuse Microwave Emission Survey (DIMES) experiment (dashed curve, from Kogut 1996; see also http://ceylon.gsfc.nasa.gov/DIMES). The (uncertain) contribution from the star–forming clouds is comparable to the galactic free-free emission and can easily be detected by the DIMES experiment. Extreme variants of our standard model (e.g., with a larger Ω_b or a lower $f_{\rm esc}$) which enhance $J_{\rm ff}$ by more than an order of magnitude, might already be ruled–out by the existing COBE limits on the spectral distortion of the CMB (Mather et al. 1994; Wright et al. 1994). The unavoidable contribution of Lyα clouds at $z \lesssim 5$ (Loeb 1996) is detectable with the future DIMES experiment, but the emission from the smooth IGM at $z \gtrsim 10$ is just below the sensitivity of this instrument in our standard model.

5.5 Conclusions

For a large range of CDM models with a Scalo IMF, the universe is reionized by a redshift $\gtrsim 10$. The optical depth to electron scattering due to reionization is in the range ~ 0.05–0.15. The resulting damping amplitude of ~ 10–20% for microwave anisotropies on angular scales $\lesssim 10°$, will be detectable with future satellite experiments such as MAP or PLANCK. These experiments might be able to isolate a damping amplitude of several percent if they measure the microwave polarization (Zaldarriaga 1996).

If the stellar IMF is strongly tilted relative to the local Scalo IMF so that most of the baryons are transformed into low mass stars, then stellar reionization is suppressed due to the absence of massive stars which ordinarily dominate the ionizing flux (cf. Fig. 5.4). In this case, however, the existence of primordial stars might be inferred from the measurable microlensing event rate that they produce in the halo of the Milky Way (Alcock et al. 1996; Flynn, Gould & Bahcall 1996) or along the lines of sight to distant quasars (Gould 1995; Dalcanton et al. 1994).

Aside from the above mentioned sensitivity to a drastic variation in the IMF, much of the qualitative results summarized in Figures 5.10–5.15 appear robust to changes in other model parameters. This robustness results from a number of reasons. The collapsed fraction depends exponentially on redshift in the rare tail of the Gaussian random field of density fluctuations–from where the first objects are drawn. The inverse dependence of the collapse redshift, and hence the scattering optical depth $\tau_{\rm es}$, on the power spectrum is therefore only logarithmic. The linear scaling of $\tau_{\rm es}$ with $\Omega_{\rm b}$ is somewhat moderated by the dependence of the recombination rate on $\Omega_{\rm b}^2$. The reionization redshift is not sensitive to variations in the star formation efficiency $f_{\rm star}$ or the escape fraction $f_{\rm esc}$, because of the strong dependence of the recombination rate on redshift ($\propto [1+z]^6$). Finally, the suppression of star–formation due to the H_2 feedback is significant compared to the suppression due to photo–ionization heating only as long as the filling factor of ionized bubbles is still small; once the ionized bubbles start to overlap, the H_2 feedback becomes unimportant.

We have calibrated the metal production in our models based on the carbon abundance detected in the Lyα forest at $z \sim 3$ (Cowie et al. 1995; Tytler et al. 1995; Songaila & Cowie 1996, Cowie 1996). If the metallicity of the Lyα forest, as inferred from the carbon/hydrogen ratio, is indeed universal and due to a single generation of stars, then the star formation efficiency for a Scalo IMF is higher than 4%, or else "hot bottom burning" (which destroys carbon in stellar envelopes) is entirely inefficient. This conclusion is independent of the details of the rest of our modeling. Future detection of other metal absorption lines in the Lyα forest would allow to discriminate among these possibilities.

Much of the uncertainty in our reionization models results from the lack of a standard theory for star formation. Existing theoretical work on star formation has been mainly guided by observations and therefore focused, by and large, on the complex environments of the interstellar medium of our galaxy. The conditions in the early universe are much better defined than those in the local interstellar medium, since the composition of the gas and hence its chemistry and cooling function are simple and well known, the power spectrum of initial density fluctuations is prescribed in specific cosmological models, and magnetic fields are probably absent before stars form. From a theoretical standpoint, the early universe offers a simpler environment for understanding star formation. The interplay between theory and observations of the reionization epoch might therefore add insight to the physics of star formation, especially if the star clusters at high redshifts are directly imaged. Our estimates, illustrated in Figure 5.16, imply that deep infrared observations with the Next Generation Space Telescope might be able to probe $\gtrsim 10^3$ bright star clusters at $z \sim 5$–10. The cumulative Bremsstrahlung emission from these clusters (cf. Figure 5.17) could potentially be detected by future instruments such as the Diffuse Microwave Emission Survey (Kogut 1996).

Chapter 6

Signatures of Intergalactic Dust From the First Supernovae

Zoltán Haiman and Abraham Loeb
Astronomy Department, Harvard University, Cambridge, MA 02138

[published in *The Astrophysical Journal*, **490**, 571 (1997)]

ABSTRACT

We quantify the consequences of intergalactic dust produced by the first Type II supernovae in the universe. The fraction of gas converted into stars is calibrated based on the observed C/H ratio in the intergalactic medium at $z = 3$, assuming a Scalo mass function for the stars. The associated dust absorbs starlight energy and emits it at longer wavelengths. For a uniform mix of metals and dust with the intergalactic gas, we find that the dust distorts the microwave background spectrum by a y–parameter in the range $(0.06\text{–}6)\times10^{-5}(M_{\rm SN}/0.3{\rm M}_\odot)$, where $M_{\rm SN}$ is the average mass of dust produced per supernova. The opacity of intergalactic dust to infrared sources at redshifts $z \gtrsim 10$ is significant, $\tau_{\rm dust} = (0.1\text{–}1) \times (M_{\rm SN}/0.3{\rm M}_\odot)$, and could be detected with the Next Generation Space Telescope. Although dust suppresses the Lyα emission from early sources, the redshifts of star clusters at $z = 10$–35 can be easily inferred from the Lyman-limit break in their infrared spectrum between 1–3.5μm.

6.1 Introduction

Recent spectroscopic observations of the Lyαforest at $z \gtrsim 3$ show evidence for a metallicity $\sim 1\% Z_\odot$ in absorbers with HI column densities as low as 10^{15} cm^{-2} (Cowie 1996; Songaila & Cowie 1996; Tytler et al. 1995). Numerical simulations identify such absorbers with mildly overdense regions in the intergalactic medium

(IGM), out of which nonlinear objects such as galaxies condense (Hellsten et al. 1997, and references therein). Indeed, damped Lyαabsorbers, which are thought to be the progenitors of present–day galaxies, show similar metalicities during the early phase of their formation at $3.5 \lesssim z \lesssim 4.5$ (Lu, Sargent, & Barlow 1996). The universality of this mean metal abundance in absorbers spanning a range of six orders of magnitude in HI column density, indicates that an early phase of metal enrichment occurred throughout the universe at $z \gtrsim 5$, before galaxies were assembled.

The formation of an early population of star clusters at redshifts $\sim$ 10–30 is a natural consequence of hierarchical structure formation in Cold Dark Matter (CDM) cosmologies (see, e.g. Haiman & Loeb 1997, hereafter HL97; Ostriker & Gnedin 1996, and references therein). The first star clusters that form in high abundance have baryonic masses $\sim 10^8 M_\odot$ and virial temperatures $\sim 10^4$ K, which allow their gas to cool via atomic transitions. Less massive objects with lower virial temperatures are unable to cool and fragment into stars due to the rapid photo-dissociation of their molecular hydrogen (Haiman, Rees, & Loeb 1997). Since the potential wells of the first clusters are relatively shallow ($\sim$10 km s^{-1}), supernova–driven winds might have expelled the metal–rich gas out of these systems and mixed it with the IGM. Incomplete mixing could have led to the observed order-of-magnitude scatter in the C/H ratio along lines-of-sight to different quasars (Rauch, Haehnelt, & Steinmetz 1997; Hellsten et al. 1997). It is an interesting coincidence that the supernova energy output associated with a metal enrichment of the IGM by $\sim 1\% Z_\odot$ corresponds to an energy release of $\sim$ 10 eV per hydrogen atom in the IGM, which is just above the binding energy of these early star clusters. Supernova feedback in these objects could have therefore determined the average metallicity observed in the Lyαforest. Direct observations of these supernovae might be feasible in the future (Miralda-Escudé & Rees 1997).

The measured C/H ratio can be used to calibrate the net fraction of gas which is converted into stars by a redshift $z \approx 3$. In HL97, we have used this fraction and an extension of the Press–Schechter formalism to calculate the early star formation history in a variety of CDM cosmologies. For a large range of models we have found that the universe is reionized by a redshift $z =$ 10–20, and that the resulting optical depth to electron scattering, ~ 0.1, is detectable with future microwave anisotropy experiments. In addition, deep imaging with future infrared telescopes, such as the Space Infrared Telescope Facility (SIRTF) or the Next Generation Space Telescope (NGST), would be able to detect pre–galactic star clusters. In particular, NGST should find $\gtrsim 10^3$ star clusters at $z > 10$ per square arcminute, given its planned detection threshold of 1 nJy in the wavelength range of 1–3.5 μm (Mather & Stockman 1996).

The early epoch of star formation and metal enrichment is inevitably accompanied by the formation of dust in supernova shells. This dust has two important observational signatures. First, the absorption of starlight energy and its re–emission at long wavelengths distorts the spectrum of the cosmic microwave background (CMB) radiation (Wright et al. 1994; Bond, Carr & Hogan 1991; Adams et al. 1989; Wright 1981). Second, the opacity of the intergalactic medium to infrared sources at redshifts $z \gtrsim 10$ could be significant. For these redshifts, infrared in the observer frame corresponds to UV in the source frame–a spectral regime in which dust absorption peaks.[1] Dust obscuration must therefore be considered when predicting the performance of future infrared telescopes such as NGST.

[1]The effect of intergalactic dust at much lower redshifts was discussed in the context of the reddening (Cheney & Rowan–Robinson 1981; Wright 1981, 1990) and the number count (Ostriker & Heisler 1984; Ostriker, Vogeley, & York 1990) of quasars.

In this paper, we quantify the magnitude of the above signatures of intergalactic dust. In § 2 we describe our method of calculation and in § 3 we compute the distortion of the microwave background spectrum and compare it to the current COBE limit. In § 4 we evaluate the opacity of the intergalactic medium to infrared sources. Finally, § 5 summarizes the implications of this work for future microwave-background and infrared observations. Throughout the paper, we adopt a Hubble constant $H_0 = 50$ km sec^{-1} Mpc^{-1} and a cosmological density parameter $\Omega = 1$.

6.2 Method of Calculation

Our first goal is to calculate the distortion of the CMB spectrum due to the production of radiation by stars and its processing through dust. The CMB spectrum is initially a pure black–body with a temperature $T_{\rm CMB} = 2.728(1+z)$ K (Fixsen et al. 1996). We assume that stars are born with the Galactic initial mass function (IMF) and use the six–segment power–law parameterization between 0.1 and 125$M_\odot$ given by Scalo (1986) (see also 5.9). We compute a composite time-dependent spectrum using standard spectral atlases (Kurucz 1993) and evolutionary tracks (Schaller et al. 1992). Our stellar modeling is described in detail in HL97. In computing the amount of dust produced by the stars, we assume that each type II supernova yields 0.3$M_\odot$ of dust which gets uniformly distributed in the intergalactic medium (see discussion in § 3), and that the dust absorption follows the wavelength-dependent opacity of Galactic dust.

To describe the effect of dust on the CMB we follow Wright (1981), and define the comoving number density of photons with a comoving frequency ν as

$$N_\nu(z) \equiv \frac{4\pi}{hc(1+z)^3} J_{\nu(1+z)}(z), \tag{6.1}$$

where $J_\nu(z)$ is the specific intensity of the background radiation field (microwave background + starlight + dust emission) in erg cm^{-2} s^{-1} Hz^{-1} sr^{-1}. The evolution of $N_\nu(z)$ is determined by the radiative transfer equation:

$$-\frac{dN_\nu(z)}{dz} = \frac{cdt}{dz}\left[j_\nu(z) - \alpha_{\nu(1+z)}(z)N_\nu(z)\right], \tag{6.2}$$

where in an $\Omega = 1$ universe, $(cdt/dz) = (c/H_0)(1+z)^{-5/2}$. Here, $j_\nu(z)$ is the comoving emission coefficient due to stars and dust,

$$j_\nu(z) = \frac{4\pi}{hc(1+z)^3} j^*_{\nu(1+z)}(z) + \frac{8\pi}{c^3}\left\{\frac{\nu^3}{\exp[h\nu(1+z)/k_{\rm B}T_{\rm dust}] - 1}\right\}\alpha_{\nu(1+z)}(z), \tag{6.3}$$

and $\alpha_\nu(z)$ is the dust absorption coefficient,

$$\alpha_\nu(z) = \rho_{\rm dust}(z)\kappa_\nu, \tag{6.4}$$

for a mass density, $\rho_{\rm dust}$, in dust. We assume that the dust is in thermal equilibrium with the total radiation field and equate the power it absorbs in the UV to the power it emits in the infrared. The dust temperature $T_{\rm dust}$ as a function of redshift is then derived from the implicit equation,

$$\int_0^\infty d\nu\,\kappa_{\nu(1+z)}\left\{N_\nu(z) - \frac{8\pi}{c^3}\frac{\nu^3}{\exp[h\nu(1+z)/k_{\rm B}T_{\rm dust}] - 1}\right\} = 0, \tag{6.5}$$

where $N_\nu(z)$ is the solution of the radiative transfer equation (6.2). We adopt the dust opacity κ_ν based on the mean Galactic extinction law $A(\lambda)/A(J)$ of Mathis (1990),

$$\kappa_\nu = 8.3 \times 10^3 \left[\frac{A(\lambda)}{A(J)}\right] \left(1 - \frac{0.084}{0.12 + (\lambda(\mu\mathrm{m}) - 0.4)^2}\right) \quad \mathrm{cm}^2\ \mathrm{g}^{-1}, \tag{6.6}$$

where $\lambda = c/\nu$ is the wavelength; the numerical coefficient converts the extinction per hydrogen atom, with the assumed gas-to-dust ratio of 100, to extinction per gram of dust; and the last term is the albedo term given by Wright (1981) for converting the extinction (absorption plus scattering) to pure absorption. We find that our results change by less than 5% if we use the sum of the theoretical graphite+silicate grain cross–sections derived by Draine and Lee (1984).

Given a specific star–formation history, stellar spectra, and dust–yields, one can integrate numerically equations (6.2)–(6.5) and get the dust temperature and the net spectrum of the radiation background at z=0. As in HL97, we express the star–formation history in terms of the mass fraction of all baryons which are assembled into collapsed objects, $F_{\mathrm{coll}}(z)$, and the fraction of these collapsed baryons which get incorporated into stars, f_{star}. The density parameter of stars is then, $\Omega_{\mathrm{star}}(z) = f_{\mathrm{star}} F_{\mathrm{coll}}(z)\Omega_{\mathrm{b}}$, where Ω_{b} is the total baryonic density parameter (= 0.05 in our standard model). The collapsed fraction $F_{\mathrm{coll}}(z)$ is calculated in a standard CDM cosmology based on an extension of the Press–Schechter formalism that takes into account pressure and photo-dissociation of H_2 (HL97, Haiman, Thoul & Loeb 1996). To study the effects of high–redshift stars alone, we do not allow stars to form at $z < 3$, i.e. we set $F_{\mathrm{coll}}(z < 3) \equiv F_{\mathrm{coll}}(z = 3)$. The contributions to the UV flux from quasars and stars at $z < 3$ that are ignored here could slightly increase the spectral distortion we derive.

The star–formation efficiency is found to be $f_{\mathrm{star}} \sim 12\%$ from the requirement that the C/H ratio be 1% of the solar value at $z = 3$. Since the collapsed fraction at $z = 3$ is $\sim$50%, the corresponding fraction of all baryons in stars is $\sim$6%. A factor of $\sim$2.3 is included in this number due to the average time required to produce carbon inside the stars, which we assume is the main–sequence time. Under this assumption only $\sim$ 43% of the total stellar carbon is produced by $z = 3$. As emphasized in HL97, carbon is mostly produced by intermediate mass ($3 - 6\mathrm{M}_\odot$) stars, whereas dust is heated primarily by the high–mass stars. Our modeling is therefore sensitive to the adopted slope of the IMF; metallicity measurements at high–redshifts for elements other than carbon would be useful in constraining the IMF at high–redshifts.

The stellar emission coefficient $j^*_\nu(z)$ is found at each redshift z as a sum of the instantaneous emission from all the stars present at z,

$$j^*_\nu(z) = f_{\mathrm{star}}\Omega_{\mathrm{b}}\rho_{\mathrm{c0}}(1+z)^3 \int_z^\infty dz' \frac{dF_{\mathrm{coll}}}{dz'} \epsilon^*[\nu(1+z'), t_{z,z'}], \tag{6.7}$$

where ρ_{c0} is the current critical density of the universe, $t_{z,z'}$ is the time interval between the redshift z' at which a star was born, and the redshift z, and $\epsilon^*(\nu, t)$ is the composite physical emissivity (in erg s^{-1} Hz^{-1} M$_\odot^{-1}$) of a cluster of stars with the Scalo IMF at time t after its initial starburst (see HL97 for details). We have assumed that all stellar photons escape their parent clouds and propagate into the intergalactic medium. In reality, a fraction of the ionizing photons will get converted into line photons with energies $E < 13.6$eV by case B recombinations inside the clouds, changing the emerging stellar spectrum. Since only a fraction $\sim 1/3$ of the stellar energy output is emitted above 13.6eV, and the dust absorption cross–sections at 13.6eV and 10.2eV differ only by $\sim 30\%$, this will only have a minor ($\sim 10\%$) effect on the total amount of energy absorbed by dust. Free-free emission by the ionized gas in the cluster can be neglected relative to the stellar emission. The mass density of dust at each redshift is given by

$$\rho_{\rm dust}(z) = f_{\rm dust} f_{\rm star} \Omega_{\rm b} \rho_{\rm c0} (1+z)^3 \int_z^\infty dz' \frac{dF_{\rm coll}}{dz'} f_{\rm dep}(t_{z,z'}), \tag{6.8}$$

where $f_{\rm dust}$ is the mass fraction of stars which gets converted into dust, and $0 \leq f_{\rm dep}(t) \leq 1$ is the fraction of the total mass in dust which is deposited after a time t following the starburst (i.e. the mass fraction of the $\geq 8{\rm M}_\odot$ stars which have completed their main sequence lifetime). Assuming that each supernova in a Scalo mass function produces $0.3{\rm M}_\odot$ of dust leads to $f_{\rm dust} = 0.001$. The uncertainties in this value will be discussed in the next section.

6.3 Distortion of the Microwave Background Spectrum due to Dust Emission

To parameterize the deviation of the spectrum at z=0 from a $T_{\rm CMB}$=2.728 K black–body, we compute the Compton y–parameter y_c, defined as

$$y_c \equiv \frac{1}{4} \left[\frac{\int d\nu N_\nu}{\int d\nu N_{0,\nu}} - 1 \right], \tag{6.9}$$

where the integrals are evaluated over the FIRAS frequency range of 60–600 GHz (Fixsen et al. 1996), and

$$N_{0,\nu} = \frac{8\pi}{c^3} \frac{\nu^3}{\exp(h\nu/k_{\rm B} T_{\rm CMB}) - 1}. \tag{6.10}$$

The y–parameter we derive is subject to a number of uncertainties associated with the formation of dust. Each of these uncertainties can be parameterized by a separate factor that multiplies our result for y_c. The first uncertainty involves the average fraction of the total stellar mass which gets converted into dust, $f_{\rm dust}$. The value $f_{\rm dust} = 0.001$ adopted in § 2 is equivalent to assuming that each star with mass above $8{\rm M}_\odot$ explodes as a supernova and produces $0.3{\rm M}_\odot$ of dust. This mass is consistent with the values inferred from observations of the depletion of the Si I line in SN 1987A (Lucy et al. 1991). Theoretical estimates with 100% condensation efficiency predict values as high as $1{\rm M}_\odot$ of dust per supernova (Dwek 1988; Weaver & Woosley 1980). If the dust yield of supernovae is supplemented by a significant amount of dust produced in stars, the value of $f_{\rm dust}$ could increase even further. On the other hand, a substantial fraction of the dust may be destroyed by sputtering inside the supernova cavity, or by shock waves from other supernovae (Dwek 1997); the y–parameter scales linearly with the survival fraction of dust, $f_{\rm surv}$. The y–parameter would also be affected by a spatial correlation between the distributions of dust and stellar sources. If dust is correlated with star forming regions, the radiation flux probed by the dust could be enhanced by a factor $f_{\rm flux}$ relative to the average flux in equation (6.7). Note that $f_{\rm flux} > 1$ as long as the dust surrounding the star forming region is optically thin; a full treatment of the three–dimensional geometry of optically thick clumps is beyond the scope of the present paper. For small values of $f_{\rm flux}$, y_c scales linearly with $f_{\rm flux}$. A related uncertainty involves the mixing efficiency of carbon with the intergalactic gas. The star–formation efficiency, $f_{\rm star}$, which is required in order to get the observed metallicity of the Lyαforest is inversely proportional to this mixing efficiency. From equations (6.2), (6.3) and (6.7) we get $y_c \propto f_{\rm star}^2$, since the radiation emitted by dust is proportional to its opacity and the stellar radiation background – both of which are independently proportional to $f_{\rm star}$. Finally, an additional uncertainty is due to the processing of the emergent stellar spectrum by neutral hydrogen, as discussed above. Depending on the amount and three-dimensional geometry of hydrogen present inside the star–forming regions, this additional processing of starlight may reduce the amount of energy absorbed by dust by a factor $f_{\rm spec}$. All of the above uncertainties can be combined through the definition of an overall efficiency parameter $\eta_{\rm y}$:

$$\eta_{\rm y} \equiv f_{\rm surv} f_{\rm flux} f_{\rm spec} \left(\frac{f_{\rm dust}}{0.001}\right)\left(\frac{f_{\rm star}}{0.13}\right)^2 \tag{6.11}$$

so that $y_c \propto \eta_{\rm y}$, and $\eta_{\rm y} \equiv 1$ in our standard model.

In addition to the uncertainties about the dust and the star–formation efficiency, we note that if an early generation of low-luminosity quasars of the type discussed by Eisenstein & Loeb (1995b) exists, these quasars could have produced UV radiation exceeding the stellar output. The UV light from these quasars would be reprocessed by dust in much the same way as the stellar radiation is, and could substantially increase the spectral distortion derived from the stars alone. The possible effects of early quasars are ignored here and will be quantified in a future paper.

Finally, there are two uncertainties related to the absorption cross section of dust: (i) the mean Galactic extinction law of Mathis (1990) is uncertain above 30μm, and (ii) theoretical modeling of the extinction law (Draine and Lee 1984) indicates that graphite and silicate dust grains contribute roughly the same amount to the Galactic extinction, with different individual cross–sections. To address these issues, we repeated our calculations (i) with the Mathis extinction law modified to $A(\lambda > 30\mu{\rm m}) \sim 1/\lambda^2$, and (ii) including graphite and silicate dust in the respective proportions 47% and 53% as found by Draine and Lee (1984), and allowing the two components independently to reach their equilibrium temperatures. In the first case, we find that changing the cross–section reduces the y–parameter by less than 2%, since the cross–section is changed only in the wavelength range that is ∼3 orders of magnitude away from the peak. In the second case, we find equilibrium temperatures of T=6.1 and 4.3K at $z = 0$ for graphite and silicate, respectively, in difference from the temperature of T=5.6K when a single temperature is assumed. However, despite this difference in temperatures, the y–parameter is reduced by less than 3% in the COBE regime of the spectrum. Thus, both of the above uncertainties have little effect on our results.

Figure 6.1 shows the full spectrum of the background radiation for our standard model, whose parameters are summarized in Table 6.1. The solid line depicts the total spectrum, which is a sum over the contributions from the original CMB, the stars, and the dust, shown separately by the dotted and dashed lines. The net contribution from dust is negative (absorption) for $\lambda \lesssim 0.015$ cm. The dust emission at longer wavelengths peaks around $\lambda \approx 0.1$ cm, within the COBE range. The y–parameter derived from this spectrum is 1.54×10^{-5}, comparable to the upper limit of 1.5×10^{-5} established by COBE (Fixsen et al. 1996).

For reference, the bottom panel of Figure 6.1 shows the star–formation history, i.e. the fraction of baryons confined in stars as a function of redshift, for our

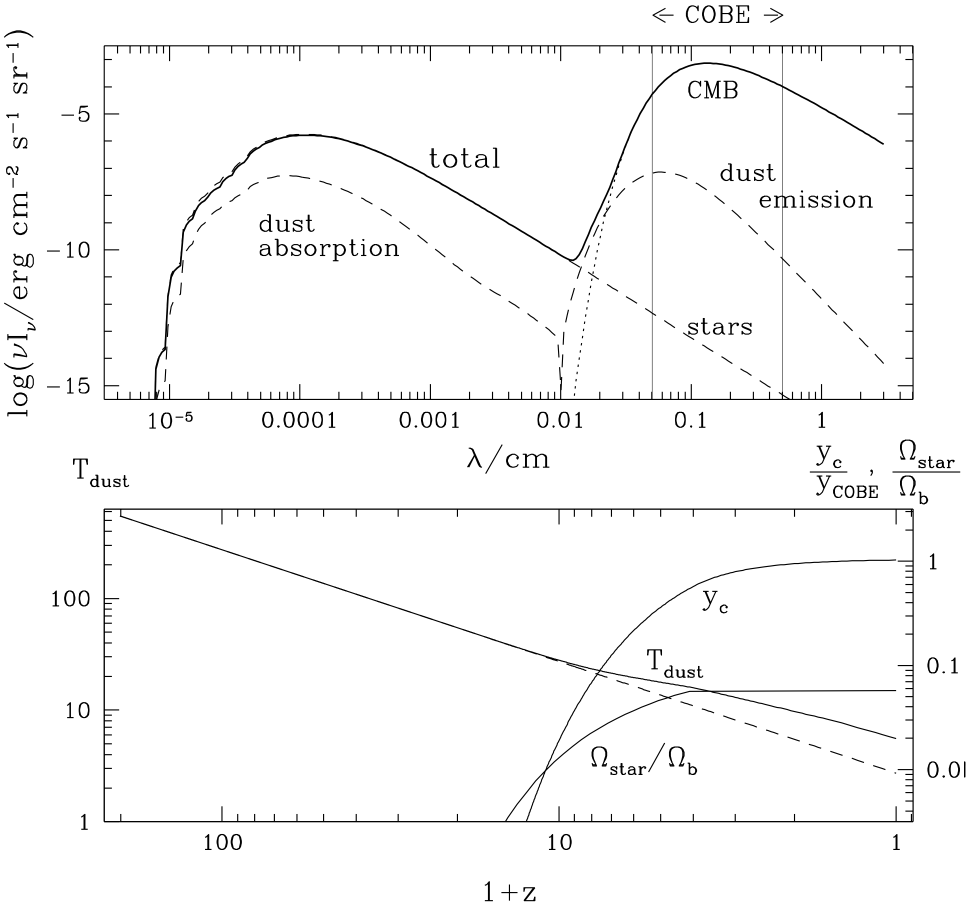


Figure 6.1: **Top panel**: *the full spectrum at $z=0$ in our standard model for the star–formation history. The solid line shows the total spectrum, the dotted and dashed lines show the separate contributions of the original CMB, the stars, and the dust. The net contribution of dust results in absorption at $\lambda \lesssim 0.015$ cm, and emission at longer wavelengths.* **Bottom panel**: *the redshift evolution of the fraction of baryons in stars, the dust temperature, and the y–parameter. Note that $f_{\rm star} \equiv F_{\rm coll}(z)^{-1}\Omega_{\rm star}/\Omega_{\rm b}$. The dashed curve shows the history of the CMB temperature.*

standard model. The same panel also shows the histories of the dust temperature and the y–parameter. The dust temperature starts to deviate noticeably from the CMB temperature (dashed line) around $z \approx 10$, when the fraction of baryons in stars $\Omega_{\rm star}/\Omega_{\rm b} = f_{\rm star}F_{\rm coll} \sim 2\%$. Note that we ignore stars and dust that form at redshifts $z \leq 3$, so that the dust emission from lower redshifts is from pre–existing dust that was heated prior to $z = 3$. Our calculation is inadequate at low redshifts, since it ignores the enhanced production of metals and dust inside galaxies and the contribution of quasars to the background radiation. The spectral distortion we obtain should therefore be regarded as a lower bound; inclusion of the contribution from lower redshifts could only increase its magnitude.

Table 6.1 shows the range of values obtained for the y–parameter as one deviates from the input parameters assumed in our standard model. In particular, if the normalization of the standard CDM power-spectrum $\sigma_{8h^{-1}}$ is increased, then star–formation starts earlier, which tends to increase the optical depth because of the

Table 6.1: *Parameter values in our standard model and its variants, and their effect on the expected values of the y–parameter and peak opacity of the intergalactic dust to a source at* $z = \infty$. *The efficiency parameters* $\eta_{\rm y}$ *and* η_τ *are defined in equations (6.11) and (6.13). Models with a predicted y–parameter in excess of the COBE upper limit,* $y_c < 1.5 \times 10^{-5}$, *are ruled out. The optical depth of intergalactic dust is significant and could affect observations with NGST.*

Parameter	Standard	Range Considered	$y_{\rm c}/10^{-5}\eta_{\rm y}$	$\tau_{\rm dust}/\eta_\tau$
$\sigma_{8h^{-1}}$	0.67	0.67–1.0	1.54–0.96	0.42–0.44
n	1.0	0.8–1.0	3.36–1.54	0.50–0.42
IMF tilt (β)	0	–0.7–1.0	0.62–1.98	0.17–0.55
H_2 feedback	yes	yes/no	1.54–0.77	0.40–0.45
$\Omega_{\rm b}$	0.05	0.01–0.1	0.06–6.16	0.08–0.84

higher density. On the other hand, the star formation efficiency in this case is reduced by 30% relative to our standard model because of the longer time available for the stars to produce carbon. The net result is an increase of the optical depth by only 5%, but a decrease of the corresponding y–parameter by $\sim 35\%$, since star–formation is now spread over a wider range of redshifts, and the average time delay between the production of radiation and dust is increased.

Tilting the power spectrum to have a lower power–law index n has the opposite effect; stars are produced later and over a narrower range of redshifts, and the star–formation efficiency has to be increased by $\sim 70\%$ because of the increased importance of the delay in the carbon production. In this case, the optical depth is increased by 20% but the y–parameter increases by more than a factor of two. Since we normalize the star–formation efficiency by fixing the total carbon yield of the stars, a change to the slope of the IMF has an effect on the net production of dust. Adding a constant $-0.7 \leq \beta \leq 1$ to the power–law index of the IMF changes the y–parameter in the range $0.62 \leq (y_c/10^{-5}\eta_{\rm y}) \leq 1.98$.

When the photo-dissociation of H_2 is ignored (e.g., due to the renewed formation of H_2 on dust grains), the threshold for fragmentation and star formation in virialized objects is lowered from virial temperatures $T_{\rm vir} \gtrsim 10^4$K (where atomic line cooling is effective) to $T_{\rm vir} \gtrsim 10^2$K (where H_2 cooling operates). The effect is similar to increasing $\sigma_{8h^{-1}}$, i.e star–formation starts earlier, but $f_{\rm star}$ is decreased by $\sim 40\%$. The net results are a slight increase in the optical depth, but a decrease in the y–parameter to $y_{\rm c}$=0.77. The input parameter that affects $y_{\rm c}$ the most is the baryon density, $y_{\rm c} \propto \Omega_{\rm b}^2$; high values of $\Omega_{\rm b}$ (such as $\Omega_{\rm b} \gtrsim 0.05$, if all other parameters are held fixed) violate the existing COBE constraint.

6.4 Optical Depth to Infrared Sources

Next, we evaluate the optical depth to absorption and scattering by dust along the line of sight to a source at a redshift $z_{\rm s} > 3$,

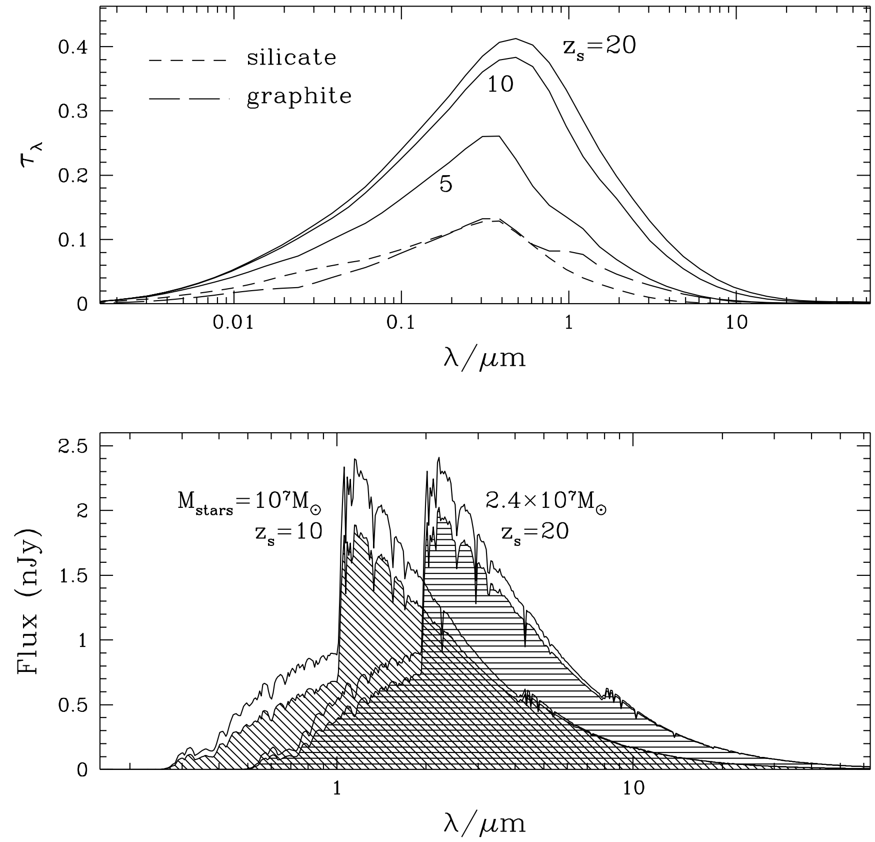


Figure 6.2: **Top panel**: *optical depth in our standard model to absorption and scattering by dust at $z \geq 3$, as a function of observed wavelength. The source is located at redshifts z_s=5, 10, or 20. The short and long dashed lines show the contributions from silicate and graphite dust grains for the $z = 5$ case.* **Bottom panel**: *the expected flux with (lower curve) and without (upper curve) intervening dust, emitted from star–clusters at z_s=10 and 20, containing 10^7 $M_\odot$ and 2.4×10^7 $M_\odot$ in stars. The stars are assumed to be distributed on the main sequence according to a Scalo IMF. The break in the spectra at the redshifted Lyman limit wavelength includes only the absorption by neutral hydrogen in the stellar atmospheres. Absorption by additional neutral hydrogen in or along the line of sight to the star cluster will modify the observed spectrum below the Lyα wavelength.*

$$\tau_{\rm dust}(\nu) = \frac{c}{H_0}\int_3^{z_s} \frac{dz}{(1+z)^{5/2}}\alpha_{\nu(1+z)}(z). \tag{6.12}$$

We ignore contributions to the integral from $z < 3$, since (i) the optical depth in homogeneously distributed dust at redshifts $z_s < 3$ would be small, and (ii) the spatial distribution of the attenuating dust at these redshifts is uncertain. Dust produced in massive galaxies with virial temperatures above 10^5K is unlikely to survive expulsion into the IGM through hot winds. Therefore, the additional dust produced at low redshifts is likely to be confined inside galaxies.

The absorption coefficient $\alpha_{\nu(1+z)}(z)$ is given in equations (6.4) and (6.6). Since we are now interested in dust obscuration along a specific line of sight, both absorption and scattering must be included, and so we omit the last term in equation (6.6). The uncertainties discussed in § 2 concerning the formation and survival of dust, as well as the fraction of the baryons converted into stars, affect the optical depth through the parameter η_τ,

$$\eta_\tau \equiv f_{\rm surv} f_{\rm flux} \left(\frac{f_{\rm dust}}{0.001}\right)\left(\frac{f_{\rm star}}{0.13}\right), \tag{6.13}$$

so that $\tau_{\rm dust}$ obtains the value predicted by our standard model times η_τ.

Figure 6.2 shows the resulting optical depth in our standard model as a function of observed wavelength λ, for three different source redshifts. The optical depth peaks at $\lambda \sim 1\mu$m, where it reaches a value $\tau_{\rm dust} = 0.4$. This high value could suppress the Lyα emission peak in the spectra of early star clusters (see Charlot & Fall 1991); attempts to detect this emission line from galaxies at $z \sim 2$–4 proved difficult in the past (Thompson, Djorgovski & Trauger 1995; Lowenthal et al. 1997). Since the total amount of dust is the same in all models with the same IMF, all other variations in the peak optical depths given in Table 6.1 arise from differences in the production history of dust. In particular, the optical depth is high when stars form early (e.g., as a result of a high $\sigma_{8h^{-1}}$ or the lack of H_2 destruction), and low when stars form late (low n). We note that the optical depths we derive are below the values that could cause a measurable reddening in the spectra of quasars at $z \lesssim 5$ (Rowan–Robinson 1995).

To illustrate the potential effect of dust on future observations with NGST, we show the composite spectrum of two star–clusters, located at $z_{\rm s}$=10 and 20, and containing 10^7 M$_\odot$ and 2.4×10^7 M$_\odot$ in stars, respectively (bottom panel). The stars are assumed to be distributed on the main sequence according to the Scalo IMF. In both cases the top curve shows the original unprocessed spectrum, and the bottom curve shows the obscured spectrum due to the intervening dust in the standard model (with $\eta_{\rm y} = 1$). These results imply that the performance of NGST near its planned detection threshold of 1 nJy in the wavelength range 1–3.5 μm, could be affected by intergalactic dust.

Another important conclusion that can be drawn from Figure 6.2 is that despite the obscuration by dust, the Lyman-limit break could easily be identified in the emission spectrum of star clusters at high-redshifts. Additional absorption by neutral hydrogen in the IGM would only make the break appear sharper. For NGST observations between 1–3.5μm, the detection of this break can identify source redshifts in the range $z_{\rm s}$ =10–35. The feasibility of this *UV dropout* technique was recently demonstrated through the successful photometric identification of galaxy redshifts in the interval $2.5 \lesssim z_{\rm s} \lesssim 3.5$ (Steidel et al. 1996; Madau et al. 1996).

6.5 Summary

We have quantified the imprint of intergalactic dust due to the first stars on the y–parameter of the CMB spectrum (in the COBE wavelength range), $y_{\rm c}$, and on the opacity to absorption and scattering of radiation from sources at high redshifts, $\tau_{\rm dust}(z_{\rm s})$. We normalized the star-formation efficiency so as to get the observed ∼1% solar C/H ratio in Lyα absorbers at z=3. With this normalization, we find $y_{\rm c} \sim 10^{-5}$ and $\tau_{\rm dust}(\infty) \sim 1$ for plausible choices of the cosmological parameters and stellar properties at high redshifts (see Table 6.1). Uncertainties due to the mixing efficiency of metals with the IGM, the stellar IMF, the destruction of dust, the spatial correlations between dust and stellar sources, or the cosmological parameters, could change these estimates by an order of magnitude. If early quasars exist, their additional UV radiation would enhance this spectral distortion.

The current COBE limit, $y_c < 1.5\times10^{-5}$, already places interesting constraints on the dust formation history of the universe at $z \gtrsim 10$. Complementary microwave anisotropy limits on the scale of $\sim 1''$ would constrain the degree of patchiness in the dust distribution. But most importantly, the future detection of source reddening at $z_s \gtrsim 10$ by NGST, would be instrumental in establishing the properties of the first stars in the universe. Despite the significant dust opacity, Figure 6.2 shows that infrared photometry by NGST could identify the redshift of sources at $10 \lesssim z_s \lesssim 35$, based on their pronounced Lyman-limit break at observed wavelengths of 1–3.5μm.

Chapter 7

Observational Signatures of the First Quasars

Zoltán Haiman and Abraham Loeb
Astronomy Department, Harvard University, Cambridge, MA 02138

[To appear in *The Astrophysical Journal*, **503**, Aug. 20 (1998)]

ABSTRACT

We study the observational signatures of a potential population of low-luminosity quasars at high redshifts in a ΛCDM cosmology. We derive the evolution of the quasar luminosity function at fainter luminosities and higher redshifts than currently detected, based on three assumptions: (1) the formation of dark–matter halos follows the Press–Schechter theory, (2) the ratio of central black hole mass to halo mass is the same for all halos, and (3) the light–curve of quasars, in Eddington units, is universal. We show that a universal light–curve provides an excellent fit to the observed quasar luminosity function at redshifts $2.6 < z < 4.5$. By extrapolating the evolution of this luminosity function to higher redshifts ($4.5 < z < 20$), we find that the associated early population of low-luminosity quasars reionizes the universe at a redshift $z \sim 12$. The reprocessing of the UV light of these quasars by dust from early type II supernovae, distorts the microwave background spectrum by a Compton y–parameter, $y \sim 10^{-5}$, comparable to the lower limit set by COBE. The Next Generation Space Telescope could detect tens of quasars from redshifts $z > 10$ per square arcminute, with its proposed 1nJy sensitivity at 1–3.5μm. Absorption spectra of several such quasars would reveal the reionization history of the universe.

7.1 Introduction

One of the outstanding problems in cosmology is the nature of the first generation of astrophysical objects which appeared when the universe first transformed from its initial smooth state to its current clumpy state. Although we have observational data on bright quasars and galaxies out to redshifts $z \sim 5$ (Schneider, Schmidt & Gunn 1991; Franx et al. 1997), and on the linear density fluctuations at redshift $z \sim 10^3$ (Bennett et al. 1996), there is currently no direct evidence as to when and how the first structures formed, and what kind of objects were responsible for end of the "dark age" of the universe (Rees 1996).

Popular Cold Dark Matter (CDM) models for structure formation predict the appearance of the first baryonic objects with masses $M \sim 10^5 M_\odot$ at redshifts as high as $z \sim 30$; objects with progressively higher masses assemble later (cf. Haiman, Thoul & Loeb 1996, and references therein). Following virialization, the gas in these objects can only continue to collapse and fragment if it can cool on a timescale shorter than the Hubble time. In the metal–poor primordial gas, the only coolants that satisfy this requirement are neutral atomic hydrogen (H) and molecular hydrogen (H_2). However, H_2 molecules are fragile, and are easily photodissociated throughout the universe by trace amounts of starlight (Haiman, Rees, & Loeb 1996, 1997; Gnedin & Ostriker 1997). Hence, most of the first stars are expected to form inside objects with virial temperatures $T_{\rm vir} \gtrsim 10^4$K. Depending on the details of their cooling and angular momentum transport, the gas in these objects is expected to either fragment into stars, or form a central black hole exhibiting quasar activity. Conversion of a small fraction ($\sim 1-10\%$) of the gas into stars or quasars could reionize the universe, and strongly affect the entropy of the intergalactic medium.

In two previous papers (Haiman & Loeb 1997a, hereafter HL97; and Loeb & Haiman, 1997, hereafter LH97), we explored the impact of the first stars on the reionization history of the universe. In this paper, we study the related signatures of early quasars, and compare their effects to those expected from the stars. For both types of sources, we calibrate the total amount of light they produce based on data from redshifts $z \lesssim 5$. The efficiency of early star formation is calibrated based on the observed metallicity of the intergalactic medium (Songaila & Cowie 1996, Tytler et al. 1995), while the early quasars are constrained to match the quasar luminosity function at redshifts $z \lesssim 5$ (Pei 1995). In HL97 and LH97 we have studied a range of standard Cold Dark Matter (CDM) cosmologies, but in this paper we focus on a particular cosmological model with a cosmological constant, namely the "concordance model" of Ostriker & Steinhardt (1995). Within this model, we predict the number of high–redshift low-luminosity quasars using the Press–Schechter formalism (Press & Schechter 1974).

We assume that the luminosity history of each black hole depends only on its mass, and postulate the existence of a a universal quasar light–curve in Eddington units. This approach is motivated by the fact that for a sufficiently high fueling rate, quasars are likely to shine at their maximum possible luminosity, which is some constant fraction of the Eddington limit, for a time which is dictated by their final mass and radiative efficiency. If the radiative efficiency were determined by a universal accretion geometry, then the bright phase of quasars would admit such a universal light–curve. We also assume that the final black hole mass is a fixed fraction of the total halo mass, and calibrate this fraction to match the typical values found in local galaxies (Kormendy et al. 1997; Magorrian et al. 1997). Based on this minimal set of assumptions we demonstrate that there exists a universal light curve, for which the Press–Schechter theory provides an excellent fit to the observed evolution of the luminosity function of bright quasars between redshifts $2.6 < z < 4.5$. Given this light–curve, we can then extrapolate the observed LF to higher redshifts and low luminosities.

Previous work on modeling the evolution of the quasar LF has been mostly phenomenological, and involved fitting ad–hoc parametric functions to the observations (Pei 1995, Boyle et al. 1991). Small & Blandford (1992) had used a universal quasar light–curve, together with the observed LF evolution, to infer the birth–rate of quasar black holes. Haehnelt & Rees (1993; see also Efstathiou & Rees 1988) adopted the Press–Schechter theory for halo formation, specified the quasar light–curve a priori, and introduced a redshift and halo–mass dependence to the black hole formation efficiency, in order to fit the observed evolution of the quasar LF. The postulated (ad–hoc) exponential decrease in the efficiency for black hole formation in small halos (with circular velocities less than 400 km/s) could be attributed to the expulsion of gas from shallow potential wells by supernova-driven winds or mergers. However, this prescription appears to be violated in nearby low-mass galaxies, such as the compact ellipticals M32 and NGC 4486B. In particular, van der Marel et al. (1997) infer a black hole mass of $\sim 3.4\times10^{6}\mathrm{M}_{\odot}$ in M32, which is a fraction $\sim 8\times10^{-3}$ of the stellar mass of the galaxy, $\sim 4\times10^{8}\mathrm{M}_{\odot}$, for a central mass-to-light ratio of $\gamma_V = 2$; while Kormendy et al. (1997) infer a black hole mass of $6\times10^{8}\mathrm{M}_{\odot}$ in NGC 4486B which is a fraction $\sim 9\%$ of the stellar mass. The efficiency of black hole formation in these systems had been comparable if not higher than in galaxies with deeper potential wells. Observations of galactic nuclei in the local universe imply black hole masses which are typically a fraction $\sim 6\times10^{-3}$ of the total baryonic mass of their host (see summary of dynamical estimates in Fig. 5b of Magorrian et al. 1997; and the maximum ratio between a quasar luminosity and its host mass found by McLeod 1997). We therefore adopt the simplest possible assumption, namely that the black hole to halo mass ratio is constant, and *infer* the quasar light–curve from observations rather than postulate it. The actual formation process of low–luminosity quasars was addressed by Eisenstein & Loeb (1995b), and Loeb (1997).

Based on the extrapolated quasar luminosity function, we quantify the effects of the first generation of low-luminosity quasars. In particular, the UV radiation produced by these quasars above the Lyman limit can lead to the reionization of the intergalactic medium (IGM), and to a number of observable signatures (Carr, Bond & Arnett 1984). First, the resulting optical depth of the universe to electron scattering at $z \lesssim 20$, $\tau_{\rm es}$, damps the microwave background anisotropies on scales $\lesssim 10^{\circ}$ (Efstathiou & Bond 1987; Kamionkowski, Spergel, & Sugiyama 1994; Hu & White 1997). The recently detected hint for a Doppler peak in ground–based microwave anisotropy experiments can already be used to set the constraint $\tau_{\rm es} \lesssim 1$ (Scott, Silk & White 1995; Bond 1995b). Future satellite experiments (such as MAP or Planck[1]) will be able to probe values of $\tau_{\rm es}$ as small as a few percent if polarization data is gathered (Zaldarriaga, Spergel, & Seljak 1997). The quasar UV flux is also absorbed by intergalactic dust, and subsequent thermal emission by this dust at longer wavelengths could introduce a substantial spectral distortion to the cosmic microwave background (CMB) radiation. Such a distortion due to population III stars have been investigated extensively in the past (Wright 1981; Adams et al. 1989; Bond, Carr & Hogan 1991; Wright et al. 1994; LH97), and its magnitude is currently constrained by COBE to have a Compton y–parameter $< 1.5\times10^{-5}$. Finally, direct detection of high–redshift low–luminosity quasars will be feasible with the Next Generation Space Telescope, whose proposed sensitivity in the 1–3.5μm range is $\sim$1nJy.

The purpose of this paper is to quantify the above signatures of high-redshift low-luminosity quasars, and to compare them with those of early stars. The paper is organized as follows. In § 2, we describe the procedure used to find the universal quasar light–curve, and the resulting fits to the observed quasar LF at $z \lesssim 5$. In

[1] See the homepages for these experiments at http://map.gsfc.nasa.gov and at http://astro.estec.esa.nl/SA-general/Projects/Cobras/cobras.html.

§ 3, we calculate the reionization history of the intergalactic medium due to early quasars, and compare it to stellar reionization. In § 4, we discuss the effects of the early quasars on the CMB, including the smoothing of its anisotropies and the distortion of its spectrum. In § 5, we predict the number counts of high–redshift quasars down to the sensitivity limit of the Next Generation Space Telescope. In § 6, we briefly discuss the sensitivity of our results to changes in the model parameters. Finally, § 7 summarizes the main conclusions and implications of this work.

7.2 Quasar Light–Curve

Our goal is to calculate the quasar luminosity function at faint magnitudes, ($\log(L_{\rm B}/\mathrm{L}_{\rm B,\odot}) \lesssim 11.5$) and at high redshifts ($z \gtrsim 4.5$), outside the regime of current observations. For this purpose, we introduce a new model which extrapolates the observed LF of bright quasars at redshifts $2.6 < z < 4.5$, based on the Press–Schechter theory and the assumption that all quasars admit the same universal light–curve in Eddington units.

The formation rate of dark–matter halos is given by the Press–Schechter formalism. We assume that black hole formation is restricted to halos whose virial temperature exceeds $T_{\rm vir} \gtrsim 10^4$K, in which atomic and Bremsstrahlung cooling allows the gas to sink to the center of the potential well. This virial temperature threshold corresponds to a minimum halo mass $M_{\rm halo} \gtrsim M_{\rm min} = 10^8 \mathrm{M}_\odot[(1+z)/10]^{-3/2}$. We further assume for simplicity that the ratio of final black hole mass to halo mass, $\epsilon \equiv M_{\rm bh}/M_{\rm halo}$ is the same for all halos. Estimates of black hole masses in nearby galaxies (Kormendy et al. 1997; Magorrian et al. 1997) imply that massive black holes weigh roughly a fixed fraction of their bulge mass, over a wide range of bulge masses $10^9\mathrm{M}_\odot \lesssim M_{\rm bulge} \lesssim 2\times10^{12}\mathrm{M}_\odot$. Since bulges are the oldest stellar components of galaxies, it is reasonable to assume that a fixed fraction of the baryonic content of early galaxies ends up in the central black hole. Note, however, that this assumption differs from the model presented by Haehnelt & Rees (1993), in which ϵ is assumed to decline exponentially for low-mass galaxies. Our simple assumption of a constant black hole mass fraction allows quasars to form only as long as the collapsed fraction of baryons is smaller than unity. This restriction can easily be remedied by a moderately nonlinear black hole formation efficiency of the form $\epsilon \propto M_{\rm halo}^{\alpha}$ with $\alpha \gtrsim 0$, which could still be consistent with the data on local galaxies. Since we focus in this paper on high redshifts where the baryonic collapsed fraction is much smaller than unity (cf. Fig. 7.7), we emphasize the simplest model in which ϵ =const. For reference, we also discuss the case $\alpha = 0.5$ in section § 6. The observational constraints from the Hubble Deep Field on the scaling of ϵ with halo mass will be examined in a future publication (Haiman, Madau & Loeb 1998).

The change in the comoving number density of black holes with masses between $M_{\rm bh}$ and $M_{\rm bh} + dM_{\rm bh}$, between redshifts z and $z + dz$, is governed by the derivative,

$$\left.\frac{d^2 N_{\rm bh}(M_{\rm bh},z)}{dM_{\rm bh}dz}\right|_{M_{\rm bh}} = \begin{cases} \frac{1}{\epsilon}\,\frac{d}{dz}\left.\frac{dN_{\rm ps}(M,z)}{dM}\right|_{M=\epsilon^{-1}M_{\rm bh}} & (\text{for } M_{\rm bh} \geq \epsilon M_{\rm min}) \\ 0 & (\text{for } M_{\rm bh} < \epsilon M_{\rm min}), \end{cases} \tag{7.1}$$

where $dN_{\rm ps}/dM$ is the Press–Schechter mass function. The actual halo formation rate is larger than the derivative $\frac{d}{dz}(dN_{\rm ps}/dM)$, since this derivative includes a negative contribution from merging halos. However, at high redshifts collapsed objects are rare, and the merger probability is low. We have compared the above expression with the more accurate result for the halo formation rate given by Sasaki (1994), and confirmed that the difference in the rates is negligible for the high redshifts and halo masses under consideration here. We therefore use equation (7.1) to describe the black hole formation rate at high redshifts.

We define $\phi(L,z)dL$ to be the number of quasars per unit comoving volume at redshift z, with intrinsic luminosities between L and $L+dL$. The luminosity function, $\phi(L,z)$, is linked to the black hole formation rate through the light–curves of the quasars. For simplicity, we make the minimal assumption that all quasars admit the same light–curve in Eddington units, i.e. their luminosity is proportional to their final mass (see §1). Under this assumption, the luminosity $L(t)$ of a quasar with a black hole mass $M_{\rm bh}$ at a time t after its birth can be written as

$$L(t) = M_{\rm bh} f(t) = \epsilon M_{\rm halo} f(t), \qquad (M_{\rm halo} > M_{\rm min}) \tag{7.2}$$

where $f(t)$ is a function of time. This leads to the expression for the LF,

$$\phi(L,z) = \int_z^\infty \int_{\epsilon M_{\rm min}}^\infty dM_{\rm bh} dz' \frac{d^2 N_{\rm bh}}{dM_{\rm bh} dz'} \delta[L - M_{\rm bh} f(t_{z,z'})], \tag{7.3}$$

where $t_{z,z'}$ is the time elapsed between the redshifts z and z'. By integrating over z', we obtain

$$\phi(L,z) = \int_{\epsilon M_{\rm min}}^\infty \frac{dM_{\rm bh}}{M_{\rm bh} \left|\dot{f}\right|_{t_\star}} \left.\frac{d^2 N_{\rm bh}}{dM_{\rm bh} dz}\right|_{z_\star} \left.\frac{dz}{dt}\right|_{z_\star}, \tag{7.4}$$

or, equivalently, by integrating over $M_{\rm bh}$ we get,

$$\phi(L,z) = \int_0^{t(z)} \frac{dt}{f(t)} \left.\frac{dz}{dt}\right|_t \left.\frac{d^2 N_{\rm bh}}{dM_{\rm bh} dz}\right|_{M_{\rm bh} = \frac{L}{f(t(z)-t)}}, \tag{7.5}$$

where $\dot{f} = df/dt$, $t_\star$ is defined by the condition $M_{\rm bh} f(t_\star) \equiv L$, and $z_\star$ is defined through the relation $t_\star \equiv t(z) - t(z_\star)$. We found it useful to utilize both of the above expressions for $\phi(L,z)$, since a comparison between them can be used to check for the numerical accuracy of the calculation. The only unknown quantities are $\dot{f}$ in equation (7.4), and f in equation (7.5). Therefore, either of these two equations can be solved to find the $f(t)$ that would fit best the observational data. Since the equations are implicit, we have found their solution iteratively. We have optimized our initial guess for $f(t)$ as follows.

The function $d^2N_{\rm bh}/dM_{\rm bh}dz$ peaks at a mass scale $M_{\rm pk} = \epsilon M_{\rm nl}(z)$, where $M_{\rm nl}(z)$ is the non–linear mass–scale at the corresponding redshift. Assuming that this function is sharply peaked around $M_{\rm pk}$, i.e. substituting $d^2N_{\rm bh}/dM_{\rm bh}dz \propto \delta(M_{\rm pk}[z])$ into equation (7.4), yields

$$\phi(L,z) \propto \frac{1}{M_{\rm pk}\left|\dot{f}\right|} \propto \frac{1}{|dL/dt|}, \tag{7.6}$$

or

$$t \propto \int_L^\infty \phi(L',z) dL'. \tag{7.7}$$

In the last step, we assumed that $dL/dt < 0$, i.e. that the quasar's luminosity decreases monotonically with time. Equation (7.7) yields the time t it takes the

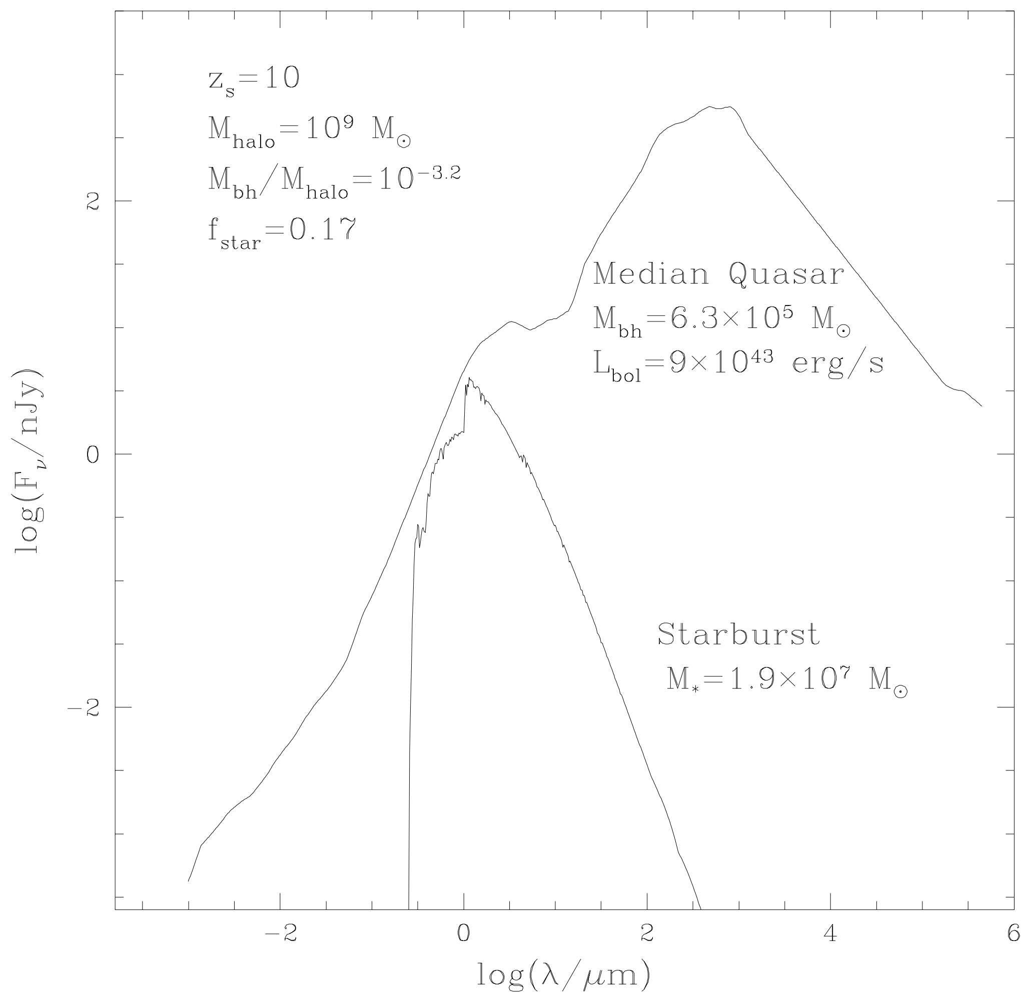


Figure 7.1: *Comparison of the observed spectral fluxes from a typical quasar and a starburst in a source located at redshift* $z_s = 10$. *The star–formation efficiency,* $f_{\rm star} = 0.17$, *was fixed so as to produce an average of 1% solar metallicity in the IGM at* $z = 3$. *The black hole formation efficiency,* $M_{\rm bh}/M_{\rm halo} = 10^{-3.2}$, *was found by our fitting procedure to the observed quasar luminosity function (see text for details).*

quasar luminosity to drop down to $L = M_{\rm pk} f$, and provides a "first guess" solution for $t(f)$ and hence for $f(t)$. The physical interpretation of the LF in this case is that it merely reflects the time spent by the quasars in each luminosity interval.

To find the light–curve $f(t)$, we need to specify $\phi(L, z)$. Given the observational data, the inferred luminosity function depends on the assumed cosmology and quasar spectrum (the latter due to k-corrections). In this paper, we adopt the "concordance model" of cosmological parameters highlighted by Ostriker & Steinhardt (1995), i.e. a flat ΛCDM model with a slightly tilted power spectrum ($\Omega_{\rm m}$, Ω_Λ, $\Omega_{\rm b}$, h, $\sigma_{8h^{-1}}, n$)=(0.35, 0.65, 0.04, 0.65, 0.87, 0.96). Convenient expressions in this model for the differential volume element, luminosity distance, and time–redshift relation were given in terms of elliptic integrals by Eisenstein (1997). An accurate arithmetic fitting formula for the growth function that appears in the Press–Schechter formula, is given by Carroll et al. (1992). For the average quasar spectrum, we have used the

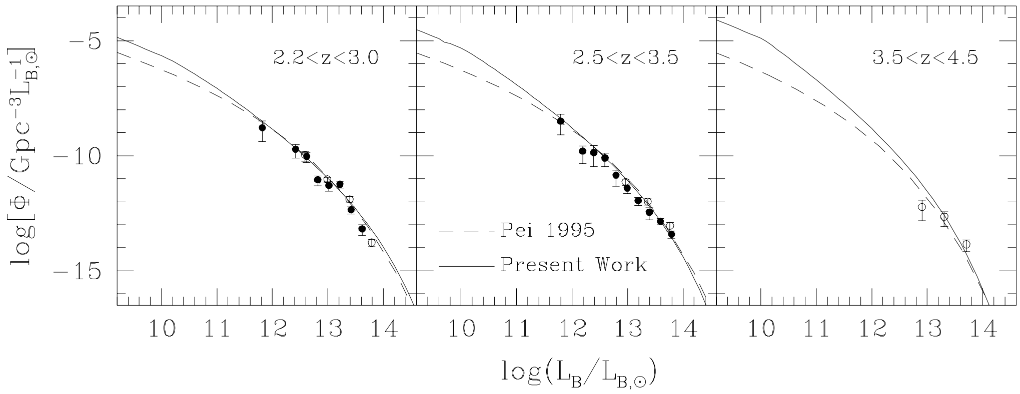


Figure 7.2: *The observed and extrapolated evolution of the quasar luminosity function in a ΛCDM cosmology. The data is taken from Pei (1995), and the dashed lines show the parametric fitting function from this reference. The solid lines show our fits based on the Press–Schechter formalism and the universal light–curve shown in Fig. 7.3.*

"median" spectrum derived from observed spectra of 47 quasars by Elvis et al. (1994). This spectrum is shown in Figure 7.1.

Pei (1995) provides a fitting formula for the B–band (0.44 μm rest–wavelength) luminosity function of observed quasars, $\phi(L_{\rm B}, z)$, including the redshift evolution both for a flat and an open $\Omega_{\rm tot} = 0.2$ cosmology. For conversion to Eddington units, we note that a $1{\rm M}_\odot$ black hole, shining at Eddington luminosity with the median flux distribution of Elvis et al. (1994), has a B–band luminosity of 5.7×10^3 times the solar B-band luminosity. Pei (1995) assumes power–law spectra for the quasars, with a slope α=0.5 in the flat model, and α=1.0 in the open model. Since near the B–band the Elvis et al. (1994) template spectrum is very close to an α=1 power–law, we used the open–universe fitting formulas from Pei (1995), with a rescaling of the luminosities and volume elements for our ΛCDM cosmology. Note that the exact conversion of the intrinsic LF to a different cosmology would require a spectrum–dependent k-correction, as well as rescaling the luminosity and the volume element, for the redshift of each individual quasar. However, we have found that using the mean redshift in each redshift bin, and assuming the α=1.0 power–law for our quasar spectrum, provided the correct LF to within several percent.

Figure 7.2 shows the resulting fitting formulas for $\phi(L_{\rm B}, z)$, together with the observational data points for the LF, in the ΛCDM cosmology at redshifts z=2.6, 3, and 4. Using the fitting formula derived by Pei (1995), equation (7.7) yields an initial guess for $f(t)$. The numerical solution for this guess can be fitted by a parameterized function, and substituted back into equation (7.4) or (7.5). The parameters of the fit can then be iteratively adjusted so as to optimize the match between the resulting $\phi(L, z)$ and the observed quasar LF at $z = 2.6, 3$, and 4. Using this procedure, we have found that an excellent fit is achieved through a simple single–parameter function for the light–curve,

$$f(t) = \frac{L_{\rm Edd}}{M_{\rm bh}} \exp\left(-\frac{t}{t_0}\right), \tag{7.8}$$

where $L_{\rm Edd} \equiv 1.4 \times 10^{38}$ erg s$^{-1}(M_{\rm bh}/M_\odot)$ is the Eddington luminosity for the final[2] black hole mass $M_{\rm bh}$. Our model, therefore, has only two free parameters, t_0 and ϵ.

[2]During the initial growth of the black hole, its mass and hence its Eddington limit, are smaller. This is not in conflict with our prescribed light–curve, as long as most of the light is emitted during a

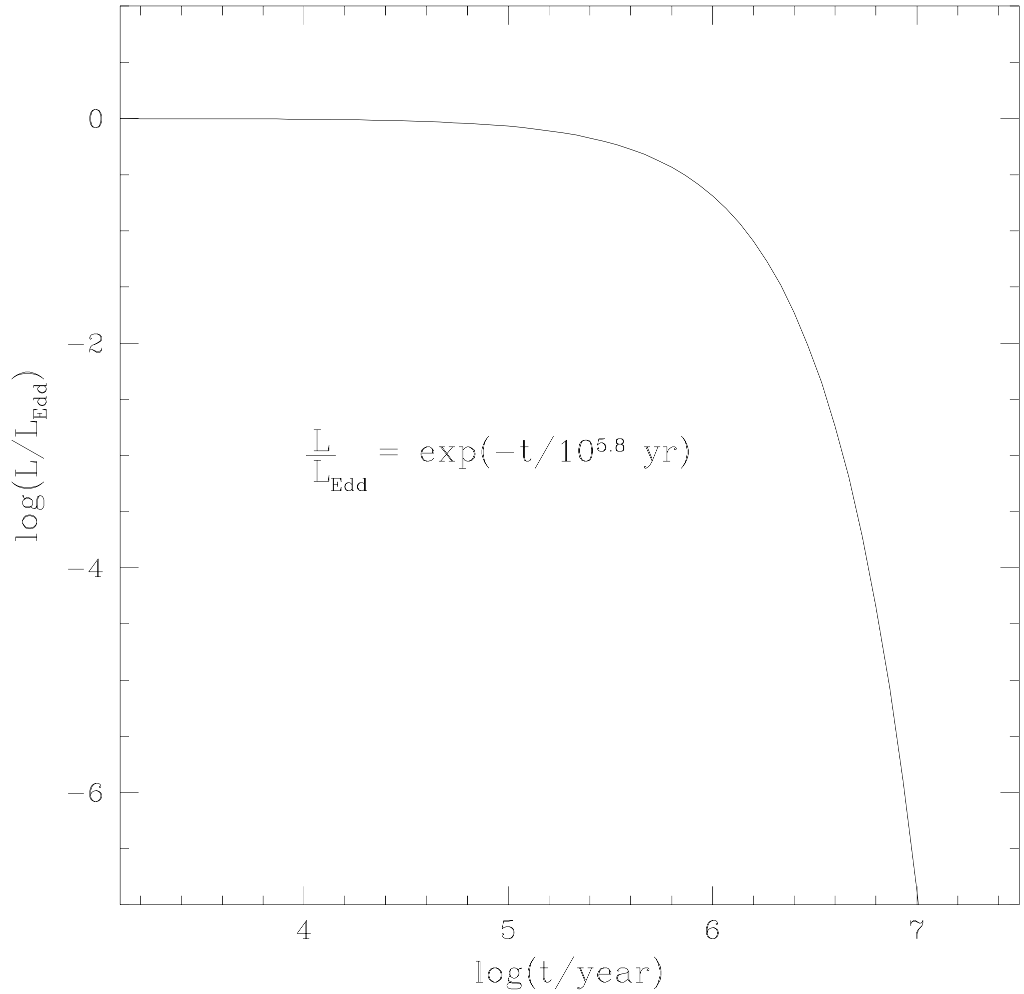


Figure 7.3: *The best–fit quasar light–curve in Eddington units, that results in the luminosity functions shown by the solid lines in Figure 7.2. The Eddington time was not used to rescale the time axis, because it does not depend on the black hole mass, and is a constant for a universal radiative efficiency.*

The best fit values of these parameters are $t_0 = 10^{5.8}$ yr and $\epsilon = 10^{-3.2}$. The best–fit light–curve is shown in Figure 7.3, while the resulting luminosity functions at $z = 2.6$, 3, and 4, are shown by the solid lines in Figure 7.2.

Our procedure results in fits to the LF comparable in quality to the original ad–hoc parametric fits given by Pei (1995). It is rather remarkable that, in the range of observed luminosities, the two curves are almost indistinguishable, even though at fainter magnitudes our curves deviate from the Pei (1995) fits, and predict many more quasars than the extrapolation of those fitting formulae. Another gratifying feature of our approach is that – with the underlying assumption that the flat portion of the light–curve corresponds to the Eddington luminosity limit of the final black hole

late phase of duration t_0, during which the black hole shines close to the Eddington limit of its final mass, $M_{\rm bh}$. The inferred low value of t_0 implies that much of the black hole mass is accumulated via accretion with a low radiative efficiency, such as predicted in models of spherical accretion (Park & Ostriker 1989) or advection–dominated accretion flows (Narayan 1996 and references therein).

– it results in a prediction for the black hole to halo mass, $M_{\rm bh}/M_{\rm halo} = 10^{-3.2}$, or black hole to gas mass $M_{\rm bh}/M_{\rm gas} = 10^{-3.2}\Omega_{\rm m}/\Omega_{\rm b} = 5.4 \times 10^{-3}$. This is close to the value derived from observational data, 6×10^{-3}, for the bulges of nearby galaxies (Kormendy et al. 1997; Magorrian et al. 1997), which might have formed at about the same time as quasars. There are, however, two shortcomings of our model. First, since it is based on the Press–Schechter formalism, it predicts that the number of bright quasars (those with masses above $M_{\rm pk}$, or $L_{\rm B} \gtrsim 10^{13}{\rm L}_{{\rm B},\odot}$) is steadily increasing down to $z = 0$. Hence, by construction, it cannot explain the decrease in the number density of bright quasars at low redshifts ($z \lesssim 2$). This decrease might be associated with the decline in merger rates at low redshifts (Carlberg 1990); or it might also be a result of the expulsion of cold gas from galaxies through supernovae-driven winds due to the peak in the star formation activity around $z \sim 2$ (Madau 1997). However, since our predictions focus on high redshifts ($z \gtrsim 5$), we ignore this shortcoming of the model. Second, the best–fit light–curve predicts a quasar lifetime of only $\sim 6.6 \times 10^5$ years, about two orders of magnitude shorter than the Eddington time at the usually assumed value of the radiative efficiency, 10%. The average radiative efficiency over the growth history of the black hole, is ~0.1%, two orders of magnitude smaller than the value for thin disks (Frank et al. 1992). This may imply that quasar black holes accumulate most of their mass in an accretion flow with a low radiative efficiency, such as advection dominated flows (Narayan 1996, and references therein). Although t_0 is an unusually short lifetime for the bright phase of quasars, it is not in conflict with existing observations. Furthermore, increasing the lifetime of quasars would enhance all of the effects that we describe below. In this sense, we are being conservative by adopting our best–fit light–curve.

The evolution predicted by equation (7.5) for the quasar LF between redshifts $2.6 < z < 18$, is shown in Figure 7.4. The break at $L_{\rm B} \sim 10^{10}{\rm L}_{{\rm B},\odot}$ is introduced by our imposed low–mass cutoff $M_{\rm halo} \geq 10^8{\rm M}_\odot[(1+z)/10]^{-3/2}$. In Figure 7.5, we show the evolution of the number density of quasars brighter than absolute blue magnitude $M_{\rm B} = -27.5$ and -22.5 ($L_{\rm B} = 10^{13.2}{\rm L}_{{\rm B},\odot}$ and $10^{11.2}{\rm L}_{{\rm B},\odot}$, respectively). The solid lines in this figure show the number densities predicted by our model, while the dotted lines show the corresponding number densities based on Pei's (1995) fitting formulas. This figure demonstrates that although our model agrees with the observational data between $2.6 < z < 4.5$, it predicts a substantially larger abundance of quasars at higher redshifts and fainter magnitudes, than the formal extrapolations of the fitting formulae given by Pei (1995).

7.3 Reionization

In this section, we derive the reionization history of the IGM in the presence of early quasars. The method of calculation follows closely the one introduced in HL97 for the case of stellar reionization. We refer the reader to this reference for more details.

Each ionizing source creates an expanding Strömgren sphere around itself. Reionization is complete when the separate HII regions overlap (Arons & Wingert 1972). Since the neutral fraction inside each HII region is small ($\sim 10^{-6}$) and the mass-weighted clumpiness of the IGM is still small at $z \gtrsim 10$ (Gnedin & Ostriker 1997), the average ionized fraction at any redshift is simply given by the filling factor of the HII regions. The reionization history depends on the formation rate of quasars, the light–curve of ionizing photons from quasars, and the recombination rate in the IGM, which are all functions of redshift.

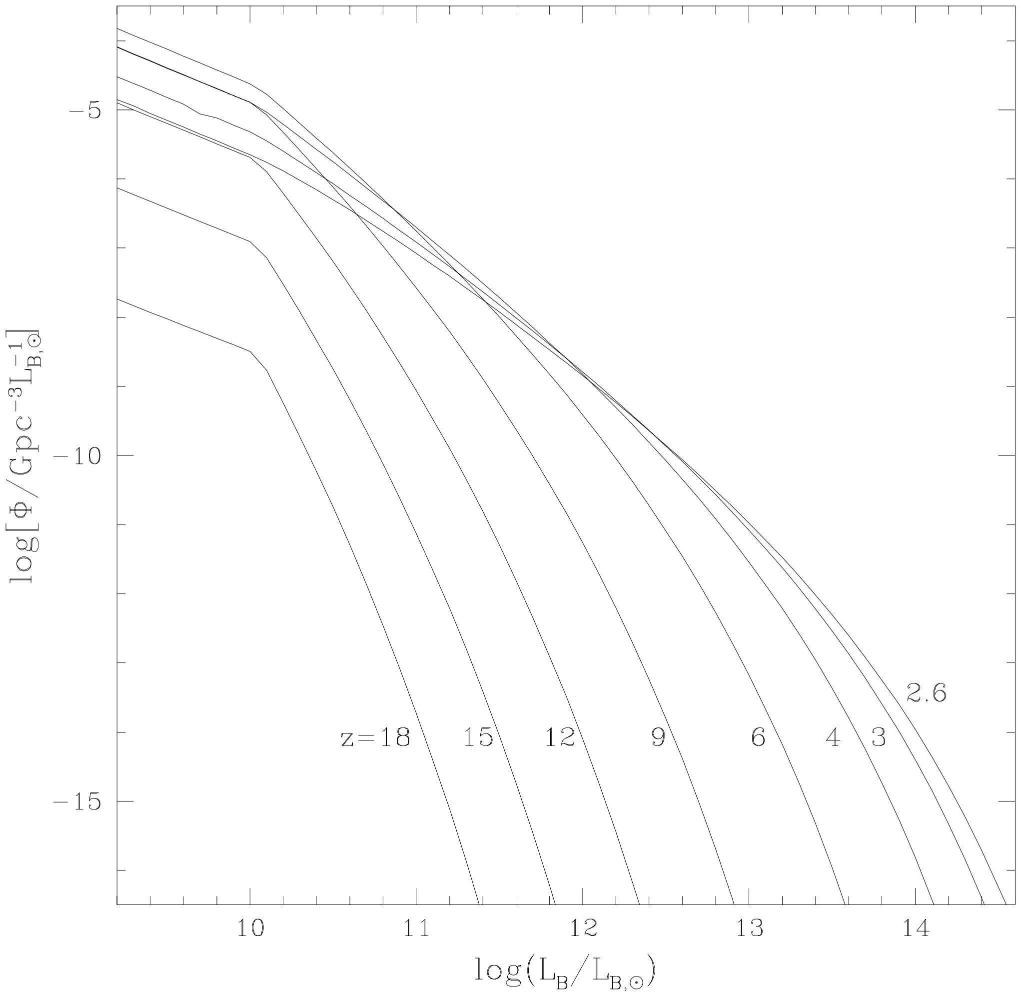


Figure 7.4: *The extrapolated quasar luminosity function at high redshifts, using the light–curve of Figure 7.3. The break at* $L_{\rm B} \sim 10^{10}{\rm L}_{{\rm B},\odot}$ *is a result of the low mass cutoff for the dark matter halo,* $M_{\rm min} = 10^8{\rm M}_\odot[(1+z)/10]^{-3/2}$*, that defines the threshold for efficient (atomic) cooling of virialized gas.*

The production rate of ionizing photons per quasar follows from the quasar spectrum and light–curve shown in Figures 7.1 and 7.3. The production rate per solar mass of the black hole, $dN_{E>13.6{\rm eV}}/dt = 6.6\times10^{47}\ M_{\rm bh}\ f(t)$ photons ${\rm M}_\odot^{-1}$ s^{-1}, is shown as the solid line in Figure 7.6. For comparison, we also show the analogous rate per stellar mass from the composite stellar spectrum in our model of a zero–metallicity starburst. Note that although the quasar rate is initially higher than the stellar rate by two orders of magnitude, it drops much more rapidly at later times. The total number of ionizing photons produced by a quasar is $\sim 1.3\times10^{61}\ M_{\rm bh}/{\rm M}_\odot$, only an order of magnitude higher than that of stars, $\sim 1.3\times10^{60}\ M_{\rm stars}/{\rm M}_\odot$.

We assume that recombinations take place only in the homogeneous IGM. This includes the implicit assumption that all of the ionizing photons escape from the neighborhood of the quasar. While this would be a poor assumption for stellar radiation, it is justified for the much more intense (and harder) quasar emission. A simple estimate can be used to show that the quasar host would be ionized shortly after the quasar turns on. Indeed, the known quasars do not show any sign of associ-

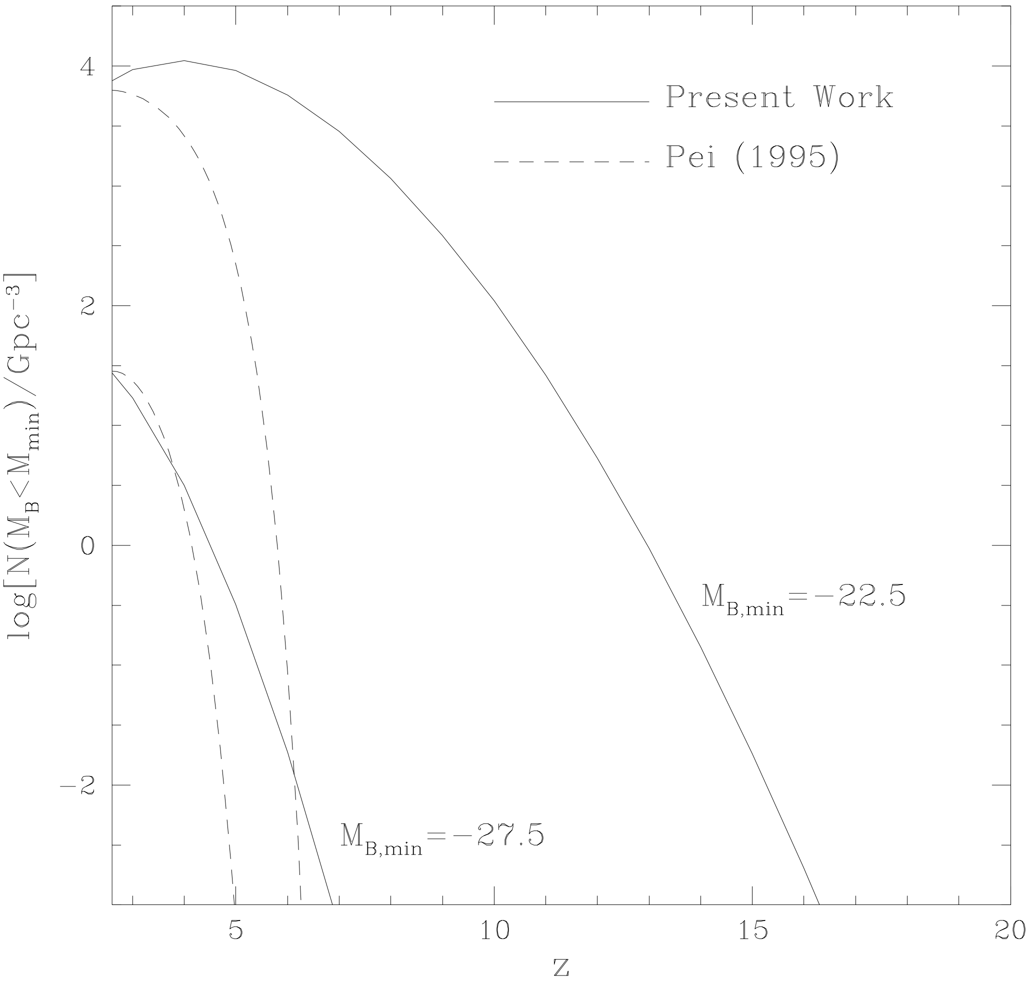


Figure 7.5: *The redshift evolution of the number density of quasars. Solid lines show the predictions from our model, while the dashed lines show the formal extrapolation of the fitting formulae obtained by Pei (1995).*

ated (galactic) HI absorption in their spectrum, while stellar spectra of high redshift galaxies often do. An additional assumption in obtaining the recombination rate is that the IGM is homogeneous, i.e. the clumping factor for ionized hydrogen is unity, $C_{\rm HII} \equiv \langle n_{\rm HII}^2 \rangle / \langle n_{\rm HII} \rangle^2 = 1$. This is justified in numerical simulations, which show that $C_{\rm HII}$ rises significantly above unity only when the collapsed fraction of baryons is large, i.e. at $z \lesssim 10$ (see Fig. 2 in Gnedin & Ostriker 1997) The evolution of a cosmological ionization front for a time–dependent source is governed by the equation,

$$n_{\rm H}\left(\frac{dr_{\rm i}}{dt} - H(t)r_{\rm i}\right) = \frac{1}{4\pi r_{\rm i}^2}\left(\frac{dN_\gamma}{dt}(t) - \frac{4}{3}\pi r_{\rm i}^3 \alpha_{\rm B} n_{\rm H}^2\right), \tag{7.9}$$

where $n_{\rm H}$ is the neutral hydrogen number density in the IGM, $r_{\rm i}$ is the physical (i.e. not comoving) radius of the ionization front, $H(t)$ is the Hubble expansion rate at cosmic time t, dN_γ/dt is the ionizing photon rate, and $\alpha_{\rm B} = 2.6 \times 10^{-13}$ cm^3 s^{-1} is the recombination coefficient of neutral hydrogen at $T = 10^4$K (see HL97, or Shapiro & Giroux 1987 for more details). The solution of this equation yields the ionized

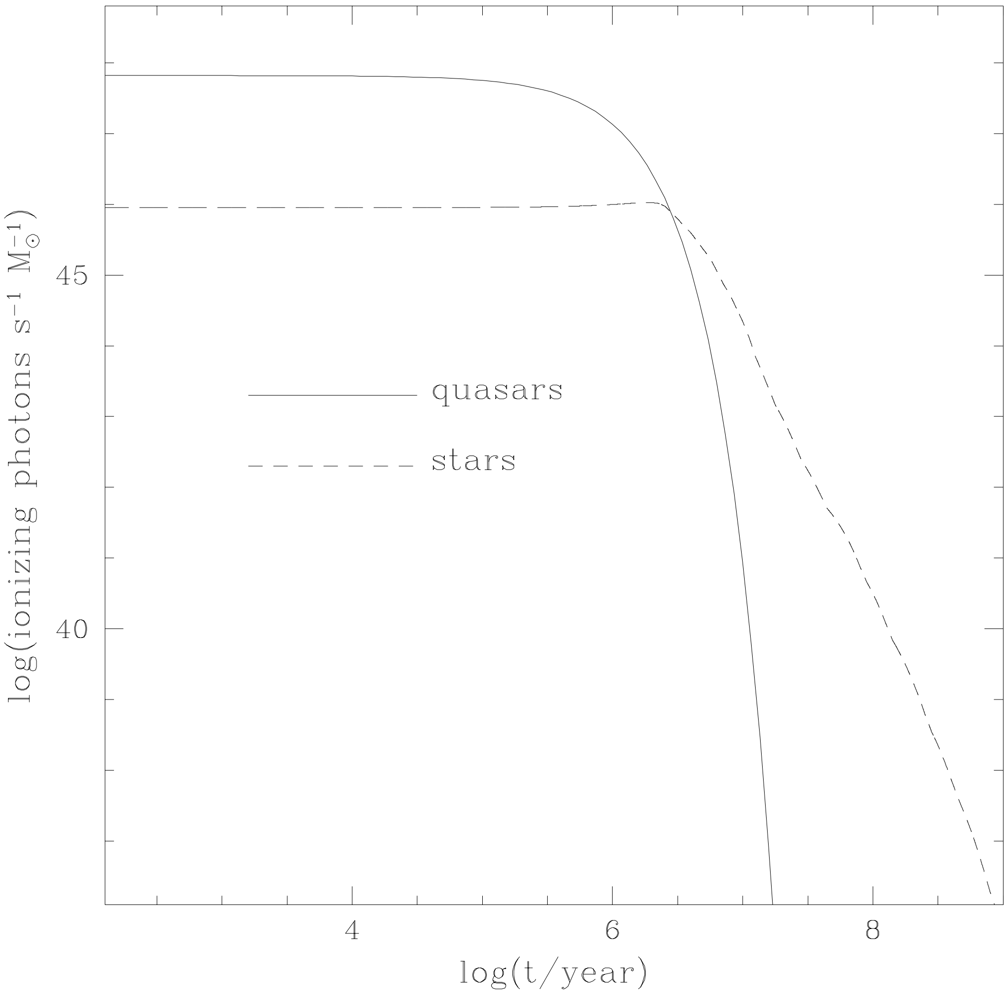


Figure 7.6: *Evolution of the production rate of ionizing photons, per unit black hole mass in quasars (solid curve), and per unit stellar mass in a starburst (dashed curve).*

volume per source, $4/3\pi r_{\rm i}^3$. A summation over all sources provides the total ionized volume within the IGM.

The results for the stellar case depend on the assumed star formation efficiency, which we normalize based on the observed metallicity of the IGM at $z \sim 3$ (HL97). In individual Lyα clouds with $N_{\rm HI} \gtrsim 10^{14}$ cm^{-2}, where C and/or Si are detected, the inferred metallicities (in solar units) are $\sim 10^{-2.5}$ for C, and $\sim 10^{-2}$ for Si (Songaila & Cowie 1996, Tytler et al. 1995). There appears to be no trend of a decreasing mean C/H or Si/H with decreasing column density, across six orders of magnitude in $N_{\rm HI}$. Hence, a plausible assumption is that the metals are universally mixed into the IGM prior to $z = 3$, so that the metallicities of the low column density absorbers are the same as that found in systems with $N_{\rm HI} \gtrsim 10^{14}$ cm^{-2}. However, as pointed out by Songaila (1997), it is also possible that there are no metals in the low column density systems; in that case, a conservative lower limit for the average metallicity is $Z_{\rm IGM} = 10^{-3} {\rm Z}_\odot$. The resulting range of IGM metallicities, $10^{-3} < (Z_{\rm IGM}/{\rm Z}_\odot) < 10^{-2}$, translates to the corresponding range $0.017 < f_{\rm star} \equiv M_{\rm star}/M_{\rm gas} < 0.17$ for the star formation efficiency.

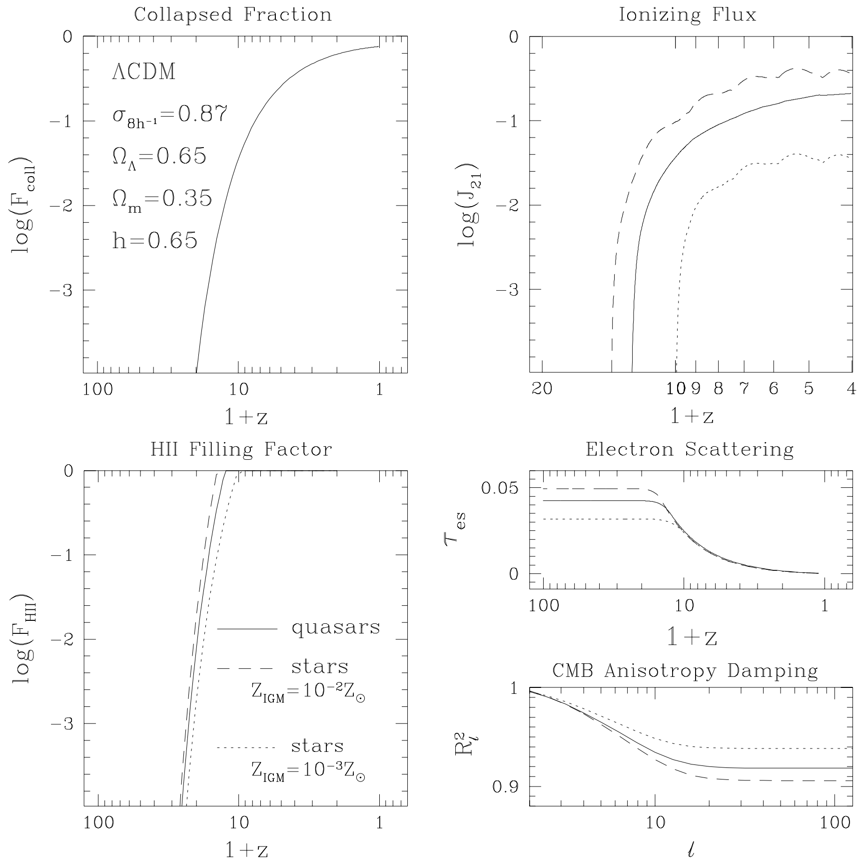


Figure 7.7: *Reionization history from quasars (solid lines) and early stars (dashed [$Z_{\rm IGM} = 10^{-2}Z_\odot$] and dotted [$Z_{\rm IGM} = 10^{-3}Z_\odot$] lines). The different panels show (i) the collapsed fraction of baryons; (ii) the evolution of the flux at the local Lyman limit frequency, in units of* 10^{-21} erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}*; (iii) the volume filling factor of ionized regions; and (iv) the optical depth to electron scattering, $\tau_{\rm es}$, and the corresponding damping factor, R_ℓ^2, for the power–spectrum decomposition of CMB anisotropies as a function of the spherical harmonic index ℓ.*

Figure 7.7 summarizes the resulting reionization histories of stars or quasars in our ΛCDM cosmology. The results for stars are shown in two cases, one with $Z_{\rm IGM} = 10^{-2}Z_\odot$ and the other with $Z_{\rm IGM} = 10^{-3}Z_\odot$, to bracket the metallicity range mentioned above. The assumed IGM metallicity does not affect the quasar calculation. The upper left panel of Figure 7.7 shows the fraction of baryons which reside in collapsed objects with a mass above $M_{\rm min} = 10^8 M_\odot[(1+z)/10]^{-3/2}$. The upper right panel shows the evolution of the average flux, J_{21}, at the local Lyman limit frequency, in units of 10^{-21} erg s^{-1} cm^{-2} Hz^{-1} sr^{-1} for quasars (solid line) and stars (dashed line for $Z_{\rm IGM} = 10^{-2}Z_\odot$; and dotted line for $Z_{\rm IGM} = 10^{-3}Z_\odot$). The equations used to obtain J_{21} are the same as described in LH97, except for the changes in the line element, time–redshift relation, and baryonic collapsed fraction, caused by our switch to a ΛCDM cosmology. The stellar mass function is taken to have the Scalo (1986) form. We have assumed that the flux between the local Lyα and Lyman–limit frequencies vanishes before reionization, due to the large Gunn–Peterson optical depth. Although this is a good approximation for the present purpose, we note that the actual spectrum in this frequency range is expected to have a sawtooth shape, as described in Haiman, Rees & Loeb (1997). After reionization J_{21} rises rapidly, and by a redshift $z = 3$ the stellar J_{21} reaches a value in the range 0.04–0.4, while quasars contribute $J_{21} \approx 0.1$. These values are consistent with the background flux level inferred from the proximity effect around quasars (Bechtold 1994).

The lower left panel of Figure 7.7 shows the resulting evolution of the ionized fraction of hydrogen, $F_{\rm HII}$. The dashed curves indicate that for the high end of the allowed metallicity range, stars ionize the IGM by a redshift $z \sim 13$; while the dotted curve shows that for the lowest metallicity, reionization is delayed until $z = 9$. (Note that for $Z_{\rm IGM} = 0.01 Z_{\odot}$, HL97 obtained stellar reionization at a redshift $z = 18$ in standard CDM. The change to $z = 13$ here results from our switch to a ΛCDM cosmology.) The solid curve shows that quasars reionize the IGM between these two redshifts, at $z = 11.5$. This result can be understood simply in terms of the total number of ionizing photons produced in each case per unit halo mass: the relative ratios of this number in the three cases are $1 \div 0.37 \div 0.1$, respectively.

Although the reionization redshift can be inferred from the corresponding damping of CMB anisotropies (see below), it is not possible to determine whether reionization was caused by stars or quasars, based on the measurement of this damping alone. However, the distinction between the two possibilities could be made if the reionization redshift of HeII is also measured. If a source at a redshift prior to reionization is discovered, the location of the Gunn–Peterson trough in its spectrum would yield the reionization redshift, while the shape of the damping wing of the trough can be used to infer the Gunn–Peterson optical depth (Miralda-Escudé 1997). Since the ionization threshold of HeII is four times higher than that of HI, the ionization of HeII requires a much harder spectrum. The relevant measure is $\eta \equiv (dN_{E>54.4\rm eV}/dt)/(dN_{E>13.6\rm eV}/dt)$, the ratio of photon fluxes above 54.4eV and 13.6eV. Equation (7.9) reveals that if recombinations are negligible, and if $\eta \leq n_{\rm He}/n_{\rm H} \approx 0.08$, then the HeII/HeIII ionization front will lag behind the HI/HII-ionization front (see Miralda-Escudé & Rees 1994, for a discussion). If recombinations are important (redshifts $z \gtrsim 15$), then the lag is increased further, since HeII recombines 5.5 times faster than HI. In our models, the HeII reionization history can be evaluated by replacing $n_{\rm H}$ with $n_{\rm He}$, and $dN_{E>13.6\rm eV}/dt$ with $dN_{E>54.4\rm eV}/dt$, and multiplying $\alpha_{\rm B}$ by a factor of 5.5, in equation (7.9).

We find that for our quasar spectrum, HI and HeII reionization occurs almost simultaneously (the HeII reionization redshift is $z = 11.8$), but for our stellar spectrum, HeII reionization is never achieved. These results are explained by fact that for the quasar spectral template of Elvis et al. (1994), $\eta = 0.09$; while for the stellar spectral template, $\eta \sim 10^{-6}$, i.e. there are almost no HeII–ionizing photons, because most of these photons are absorbed already in the stellar atmospheres (see Fig. 5.3). It therefore seems likely that intergalactic HeII was reionized by quasars rather than by stars. Thus, the reionization history of HeII provides an ideal probe for the formation history of the first quasars.

Our results indicate that the reionization redshifts of HI and HeII should not be very different, unless we have substantially overestimated the emission from high–redshift quasars, or underestimated the abundance of high mass stars. Reimers et al. (1997) have recently claimed a detection of the HeII reionization epoch at $z = 3$. Their claim is based on the lack of any detected flux above the redshifted HeII Lyα frequency, although the flux at the corresponding redshifted HI Lyα frequencies is transmitted without any absorption. This observation, however, only places the limit $\tau_{\rm He} \gtrsim$ few on the HeII Gunn–Peterson optical depth, rather than the $\tau_{\rm He} \gtrsim$ few thousand, required to prove HeII reionization. In fact, Miralda–Escudé (1997) argued that these observational results are likely to have been caused by spatial variations in the gas density, and fluctuations of the background radiation intensity above the HeII edge, rather than the fact that the HeIII bubbles have not yet overlapped. In summary, the interpretation of this, and other recent detections of HeII absorption (Jacobsen et al. 1994; Tytler et al. 1995; Davidsen et al. 1996; Hogan et al 1997) are still controversial, in that these measurements are consistent with the absorption arising either from non-overlapping HeIII bubbles, or from other effects associated with intervening discrete systems (see, e.g. Madau & Meiksin 1994; Giroux & Shull 1997).

7.4 Signatures Imprinted on the Cosmic Microwave Background

Reionization results in a reduction of the temperature anisotropies of the cosmic microwave background (CMB). The free electrons released during this epoch increase the optical depth of the universe to electron scattering ($\tau_{\rm es}$), and damp the CMB anisotropies on scales below the size of the horizon at that time, $\lesssim 10°$ (Efstathiou & Bond 1987; Kamionkowski, Spergel, & Sugiyama 1994). The electron scattering optical depth in our models is approximately $\tau_{\rm es}(z) \approx 0.053\Omega_{\rm b}h \int dz\sqrt{1+z}F_{\rm HII}(z)$, where $F_{\rm HII}(z)$ is the filling factor of the HII regions. Hu & White (1997) provide a fitting formula for the damping factor of the CMB power–spectrum, R_ℓ^2, as a function of the index ℓ in the spherical harmonic decomposition of the microwave sky [the angular scale corresponding to a given ℓ–mode is $\sim 1° \times (\ell/200)^{-1}$]. The damping factor is uniquely related to the evolution of the optical depth for electron scattering as a function of redshift.

The lower right panel of Figure 7.7 shows both $\tau_{\rm es}(z)$ and R_ℓ^2 for the stellar and quasar reionization histories. The electron scattering optical depth is 3–5% for the stellar case, and ~4% for the quasar case. The corresponding damping factors for the power spectrum of CMB anisotropies are 6–10%, and ~8%, respectively. Although the amplitude of these damping factors is small, they are within the proposed sensitivities of the future MAP and Planck satellite experiments, if data on both temperature and polarization anisotropies of the CMB will be gathered (see Table 2 in Zaldarriaga et al. 1997).

The UV emission from early quasars could also distort the spectrum of the CMB, after being reprocessed by dust. An early epoch of star formation and metal enrichment is inevitably accompanied by the formation of dust in supernova shells. This dust would absorb the UV flux from both stars and quasars, and re–emit it at longer wavelengths. The re–emitted radiation is added to the CMB spectrum, and introduces a deviation from its perfect blackbody shape. A distortion of this type due to population III stars alone has been considered by several authors (Wright et al. 1994; Bond, Carr & Hogan 1991; Adams et al. 1989; Wright 1981).

In a previous paper (LH97), we have calculated the distortion amplitude from early stars in a standard CDM cosmology, using the Press–Schechter formalism and normalizing the star formation efficiency in collapsed objects based on the observed metallicity of the IGM. The essential assumptions in this calculation were: (1) the initial mass function (IMF) of stars at high–redshifts is the same as observed locally (Scalo 1986), and (2) each type II supernova yields $0.3{\rm M}_\odot$ of dust, which gets uniformly distributed within the intergalactic medium. We further assumed that the dust is in thermal equilibrium with the total (CMB + stellar) radiation field, that its

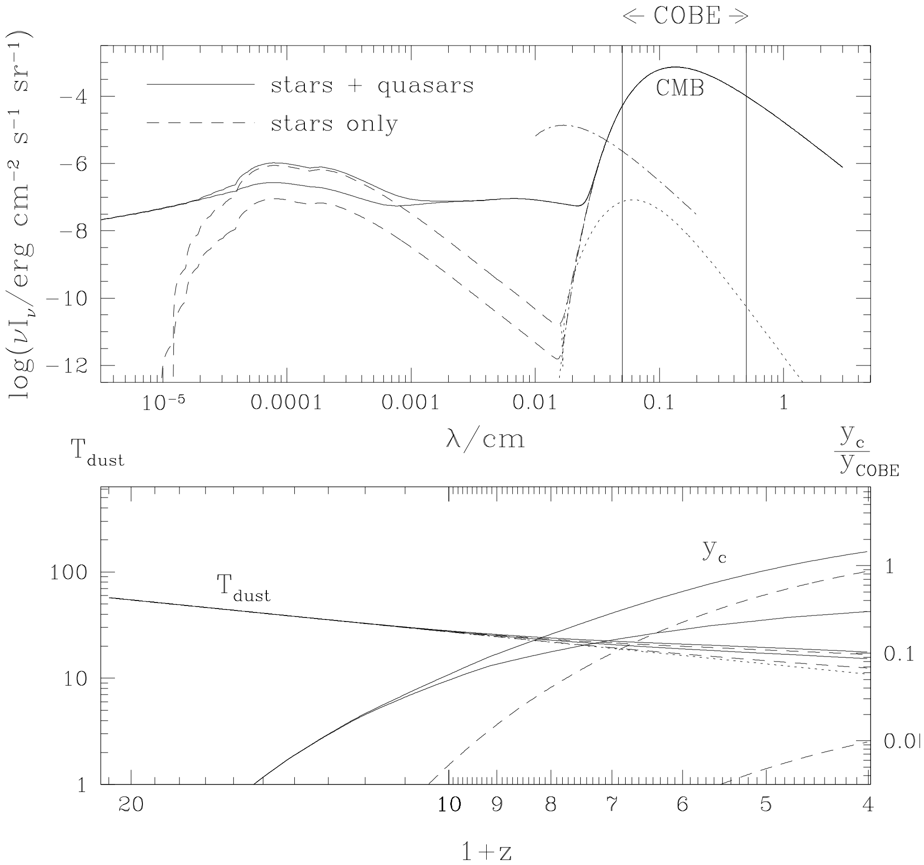


Figure 7.8: *Upper panel: The full (CMB + direct light + dust emission) background flux at redshift $z = 3$ in the presence of early stars alone (dashed lines) and in the presence of both stars and quasars (solid lines). The pairs of dashed and solid lines bracket the possible range of the IGM metallicity, $10^{-3}Z_\odot < Z_{\rm IGM} < 10^{-2}Z_\odot$. The dotted line shows the dust emission alone in the model with both stars ($Z_{\rm IGM} = 10^{-2}Z_\odot$) and quasars. The dot–dashed line shows the recent COBE detection of the extragalactic far infrared background at $z = 0$ (Fixsen et al. 1998). Lower panel: The corresponding redshift evolution of the dust temperature, and the Compton y–parameter. The dotted curve shows the CMB temperature, $T_{\rm CMB} = 2.728(1 + z)K$.*

absorption follows the wavelength-dependent opacity of Galactic dust (Mathis 1990), and that it radiates as a blackbody at its equilibrium temperature. We have found that allowing two dust components made of graphite and silicate (Draine & Lee 1984) with independent equilibrium temperatures, does not change our results by more than a few percent. For a range of possible parameter values within standard CDM, we found that the opacity of intergalactic dust to infrared sources at redshifts $z \gtrsim 10$ is significant, $\tau_{\rm dust} = (0.1–1)$, and that this dust distorts the microwave background spectrum by a Compton y–parameter in the range $(0.06–6) \times 10^{-5}$.

Here, we repeat the LH97 calculations in a ΛCDM cosmology, and extend them by adding the UV flux from quasars to the total radiation field. Our new results are presented in Figure 7.8. The top panel of this figure shows the resulting total spectrum of the radiation background (CMB + direct stellar and quasar emission + dust emission) at $z = 3$. The spectrum describes a minimum level of CMB distortion down to $z = 3$; more distortion could be added between $0 < z < 3$ by dust and radiation from galaxies. However, the additional emission at $0 < z < 3$ does not relate directly to the reionization epoch and is therefore ignored here. The dashed lines show the spectrum for the stellar case, for $Z_{\rm IGM} = 10^{-2}{\rm Z}_\odot$ (top curve) and $Z_{\rm IGM} = 10^{-3}{\rm Z}_\odot$ (bottom curve). The pronounced peak at a wavelength of $\sim 1\mu$m is the sum of the direct starlight from high–z stars, redshifted to $z = 3$. The deviation from the pure $T_{\rm CMB} = 2.728(1 + z)$K blackbody shape (shown by the dotted curve) are visible in this figure below a wavelength of 0.02cm. As an example, the dotted line shows the dust emission explicitly in the model when radiation is added both from stars (with $Z_{\rm IGM} = 10^{-2}{\rm Z}_\odot$) and quasars. For comparison, we also show the recent FIRAS detection of the extragalactic far infrared background at $z = 0$ (dot–dashed line, from Fixsen et al. 1998), which exceeds our prediction for the high redshift dust emission, as it should. The dust distortion in our models in the FIRAS range can be quantified by the Compton y–parameter, defined as

$$y_{\rm c} \equiv \frac{1}{4}\left[\frac{\int d\nu N_\nu}{\int d\nu N_{0,\nu}} - 1\right], \tag{7.10}$$

where the integrals are evaluated over the FIRAS frequency range of 60–600 GHz on board the COBE satellite (Fixsen et al. 1996), N_ν is the comoving number density of photons with a comoving frequency ν in the total radiation field, and $N_{0,\nu}$ reflects the unperturbed CMB,

$$N_{0,\nu} \equiv \frac{8\pi}{c^3}\frac{\nu^3}{\exp(h\nu/k_{\rm B}T_{\rm CMB}) - 1}. \tag{7.11}$$

The bottom panel of Figure 7.8 shows the redshift evolution of the y–parameter, as well as the dust temperature, $T_{\rm dust}$. For the stellar cases considered above, we obtain $1.3\times10^{-7} < y_{\rm c} < 1.2\times10^{-5}$ at $z = 3$ (dashed lines). Since both the dust opacity and the stellar radiation background are independently proportional to the star–formation efficiency $f_{\rm star}$, the spectral distortion is proportional to $f_{\rm star}^2$ (see LH97). For the high end of the allowed metallicity range, the distortion at $z = 3$ is just below the upper limit set by COBE, $y = 1.5 \times 10^{-5}$ (Fixsen et al. 1996); for lower metallicities, the y–parameter drops rapidly below this value. The dust temperature remains in all cases close the the CMB temperature (shown by the dotted line).

The solid lines in Figure 7.8 show how these results change when the radiation from the early quasars is added. In particular, the top panel shows that the broad emission spectrum of quasars (cf. Fig. 7.1) adds a background radiation field over a much broader range of wavelengths than stellar emission. But since dust is effective in absorbing the quasar flux only in a relatively narrow range around $\sim 1\mu$m, the distortion in the COBE regime is not significantly altered. In particular, we find $4.1 \times 10^{-6} < y_{\rm c} < 2.0 \times 10^{-5}$ at $z = 3$ for the range $10^{-3}{\rm Z}_\odot < Z_{\rm IGM} < 10^{-3}{\rm Z}_\odot$. It is interesting to note that the long-wavelength emission from high redshift quasars in the COBE wavelength range, results in a minimum distortion of $y_{\rm c} \approx 3.4 \times 10^{-6}$. Depending on $Z_{\rm IGM}$, this amounts to a fraction $\sim$ 17–82% of the total y–parameter values quoted before. This minimal level of distortion results simply from the cumulative far-infrared flux of early quasars, and should be present even in the absence of any intergalactic dust. It can be avoided only if early quasars have much less emission longward of $\sim 100\mu$m, relative to the spectrum shown in Figure 7.1.

7.5 Number Counts of Faint Quasars

Next, we examine the feasibility of direct detection of the early population of low-luminosity quasars. Although any such detection is beyond the sensitivities of current observing programs, future instruments such as the Space Infrared Telescope Facility (SIRTF) or the Next Generation Space Telescope (NGST), will achieve the required sensitivity. NGST, which is scheduled for launch a decade from now, is expected to reach a sensitivity of ~ 1 nJy in the wavelength range 1–3.5μm (Mather & Stockman 1996; see also http://ngst.gsfc.nasa.gov). The number of sources that NGST would detect at redshifts $z > z_{\rm min}$, per unit solid angle, with observed flux between F_ν and $F_\nu + dF_\nu$ (averaged over the observed wavelength range 1–3.5μm), is given by

$$\frac{dN}{d\Omega dF_\nu}(F_\nu, z_{\rm min}) = \int_{z_{\rm min}}^{\infty} dz \left(\frac{dV_{\rm c}}{dz d\Omega}\right) n_{\rm c}(z, F_\nu), \tag{7.12}$$

where $dV_{\rm c}/dz d\Omega$ is the comoving volume element per unit redshift per unit solid angle, and $n_{\rm c}(z, F_\nu)$ is the comoving number density of objects at redshift z, whose flux at $z = 0$ is observed to be between F_ν and $F_\nu + dF_\nu$. Based on the halo formation rate $d^2N_{\rm ps}/dMdz$ [cf. Eq. (7.1)], $n_{\rm c}$ is given by a sum over halos of different ages that exist at each redshift,

$$n_{\rm c}(z, F_\nu) = \int_{z}^{\infty} dz' \frac{dM}{dF_\nu}(z, z', F_\nu) \left.\frac{d^2N_{\rm ps}}{dMdz'}\right|_{M(z,z',F_\nu),z'}, \tag{7.13}$$

where the factor dM/dF_ν converts the number density per unit mass interval to number density per unit flux interval, based on equation (7.14). The assumptions implicit in this expression are that the simple derivative $d^2N_{\rm ps}/dMdz$ gives the halo formation rate [see discussion following eq. (7.1)], and that the halos do not grow in mass during the active phase of the emission sources they host (a good assumption for the short–lived quasar activity). Relaxing any of these assumptions would increase our predicted number counts. Note that, in HL97 we had effectively identified $z = z'$ in the kernel of equation (7.13), when estimating the number counts of star clusters. This might be justified only for stellar sources that live for a Hubble time or longer. However, in this case there is a danger of double counting merging halos in neighboring redshift bins (although at sufficiently high redshifts objects are exponentially rare and their merging probability is small). We intend to address this issue more rigorously using the excursion-set formalism (Bond et al. 1991; Lacey & Cole 1993, 1994), in a future publication.

The average flux $F_\nu(z, M)$ from a halo at redshift z with mass M, that was formed at $z' \geq z$ equals

$$F_\nu(z, z') = \frac{10^{32}}{d_{\rm L}(z)^2} \frac{\epsilon M}{\Delta\nu_0} \int_{\nu_{\rm min}}^{\nu_{\rm max}} d\nu j(\nu, t_{z,z'}) \qquad \text{nJy}. \tag{7.14}$$

Here the halo mass M is in units of $M_\odot$; $\nu_{\rm min}$ and $\nu_{\rm max}$ are the redshifted frequencies (in units of Hz) corresponding to the observed wavelengths of 3.5μm and 1μm, respectively; $\Delta\nu_0 = 2.14 \times 10^{14}$ Hz is the width of the NGST frequency band; $d_{\rm L}(z)$ is the luminosity distance in cm; $\epsilon = 10^{-3.2}$ for quasars or $\epsilon = 0.17\Omega_{\rm b}/\Omega_{\rm m} = 0.02$ for stars, is the fraction of halo mass in the form of the central black hole or stars, respectively; $j(\nu, t_{z,z'})$ is the template luminosity per unit (stellar or black hole) mass for either stars or quasars, in units of erg sec^{-1} Hz^{-1} sr^{-1} $M_\odot^{-1}$; and $t_{z,z'}$ is the time

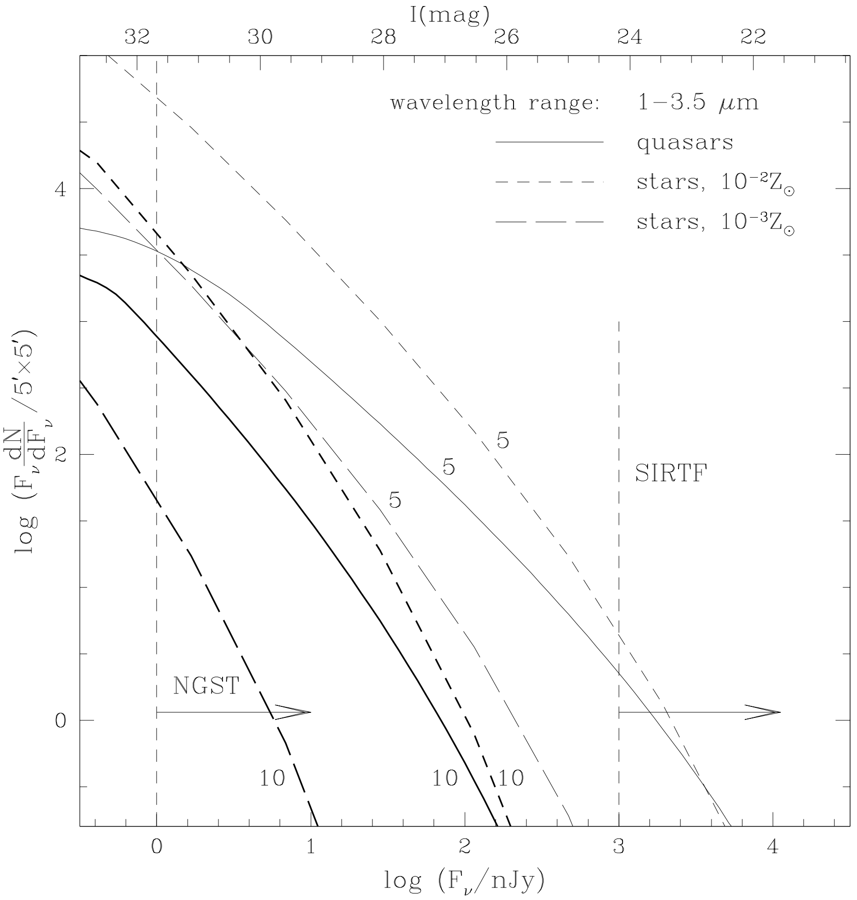


Figure 7.9: *The number per logarithmic flux interval of high–redshift objects which could be probed by future space telescopes, in the wavelength range of 1–3.5μm. We assume sudden reionization at z=11.5 for quasars, and at z=9 or 13 for stars. The vertical dashed lines show the expected sensitivities of the Space Infrared Telescope Facility (SIRTF) and the Next Generation Space Telescope (NGST). The thick lines, labeled "10", correspond to objects located at redshifts* $z > 10$, *and the thin lines, labeled "5", correspond to objects located at redshifts* $z > 5$. *The upper labels on the x–axis are Johnson I magnitudes.*

elapsed between redshifts z' and z. The time dependence of $j(\nu)$ is obtained from our fitted light–curve function $f(t)$ in equation (7.2) for quasars, and from tabulated evolutionary tracks for stars (Schaller et al. 1992; see HL97 for details). We suppressed the emergent flux from all objects at emission frequencies above Lyα before reionization occurs (assuming sudden reionization at the redshift $z = 13$ or 9 for stars, and $z = 11.5$ for quasars), since the Gunn–Peterson optical depth of HI at these frequencies is exceedingly high prior to reionization.

Figure 7.9 shows the predicted number counts normalized to the field of view of SIRTF (which is scheduled for launch long before NGST), $5' \times 5'$; the planned field-of-view of NGST is similar, $4' \times 4'$. Figure 7.9 shows separately the number per logarithmic flux interval of all objects with $z > 5$ (thin lines), and with $z > 10$ (thick lines). In general, the number of detectable sources is high. With its expected sensitivity, NGST will be able to probe about 10^3 quasars at $z > 10$, and $10^{3.5}$ quasars at $z > 5$ (solid curves) per field of view. The bright–end tail of the number counts approximately follows the power law $dN/dF_\nu \propto F_\nu^{-2.5}$. This slope is too shallow for foreground galaxies to significantly increase the number counts due to the amplification bias from gravitational lensing. The top solid curve also shows that the number of faint quasars starts to turn over below ~ 1 nJy, as a result of the cutoff in the minimum halo mass that allows atomic line cooling. The average angular separation between the quasars at $z > 10$ would be $\sim 10''$, well above the angular resolution limit of $0.06''$ planned for NGST. For comparison, we show in Figure 7.9 the corresponding number counts of "star–clusters", i.e. assuming that each halo shines due to a starburst that converts a fraction 0.017-0.17 of the gas into stars. The dashed lines indicate that NGST might detect $10^{1.5}$–$10^{3.5}$ star–clusters at $z > 10$ per field of view, and $10^{3.5}$–$10^{4.5}$ clusters at $z > 5$. Unlike quasars, star clusters could in principle be resolved, if they extend over a scale comparable to the virial radius of the collapsed halo (Haiman & Loeb 1998a).

7.6 Sensitivity to Input Parameters

Our extrapolation of the quasar LF to high redshifts and low luminosities depends on the fitting procedure described in § 2, with the resulting best–fit LFs shown in Figures 7.2 and 7.4. This fit has two free parameters, namely the black–hole to halo mass ratio, ϵ, and the lifetime of the black–hole, t_0 [the latter is related to the radiative efficiency of the accreting gas $\varepsilon_{\rm rad}$ by $\varepsilon_{\rm rad} = {\rm L_{Edd}} t_0/{\rm M_\odot} c^2 = 1.5 \times 10^{-3}(t_0/10^{5.8}{\rm yr})$]. Figure 7.10 shows how the LF changes at three different redshifts, $z = 2.6$, 3, and 4, when ϵ and t_0 are varied independently. The thick curves depict the best–fit LF ($\epsilon = 10^{-3.2}$ and $t_0 = 10^{5.8}$ years), and the left column shows the LFs when ϵ is increased by a factor of 1.5, 2, or 4 (upper curves), or decreased by the same factors (lower curves). Similarly, the right column shows the the LFs when t_0 is increased (upper curves) or decreased (lower curves) by these amounts.

Figure 7.10 provides an estimate for the tolerance of our fit. The LF is more sensitive to changes in ϵ than to changes in t_0: changing ϵ by a factor of 1.5 or t_0 by a factor of 2 visibly decreases the quality of the fit. We have found that this result still holds when both parameters are varied simultaneously: if ϵ or t_0 is changed by a factor of 1.5 or 2 respectively, the other parameter cannot be adjusted to recover the quality of the original fit. This results from the fact that changing ϵ tilts the LF, while changing t_0 is almost equivalent to a simple multiplicative scaling. In summary, given the quasar LF data at three different redshifts, our fitting procedure determines ϵ to $\sim$50% and t_0 to roughly a factor of two. The allowed range of these two parameters translates to an allowed range in our results. As an example, a factor of 1.5 increase/decrease in ϵ causes the derived quasar reionization redshift $z_{\rm ion}$, to vary between 12.3 and 11, while a factor of 2 increase/decrease in t_0 results in $12.9 > z_{\rm ion} > 10.3$.

The third important parameter in our models is the minimum halo mass $M_{\rm min}$. Although it does not effect the LF in the observed range ($\log(L_{\rm B}/{\rm L_{B,\odot}}) \gtrsim 11.5$), this parameter does effect the LF at faint luminosities ($\log(L_{\rm B}/{\rm L_{B,\odot}}) \lesssim 10$, cf. Figure 7.4), and it therefore effects the reionization redshift. In our model, we have chosen the cutoff mass $M_{\rm min} = 10^8 {\rm M_\odot}[(1+z)/10]^{-3/2}$, based on the requirement of efficient cooling by atomic H (assuming that H_2 is photo-dissociated by the UV background below the Lyman limit, cf. Haiman, Rees & Loeb 1997). If molecular hydrogen (H_2) was still an effective coolant at high redshifts, then the expected minimum mass would be lowered by ~ 3 orders of magnitude. For the quasar case, the presence of additional $\sim 10^5 - 10^8 {\rm M_\odot}$ halos would increase $z_{\rm ion}$ from 11.5 to 18.5, with a corresponding $\sim$90%

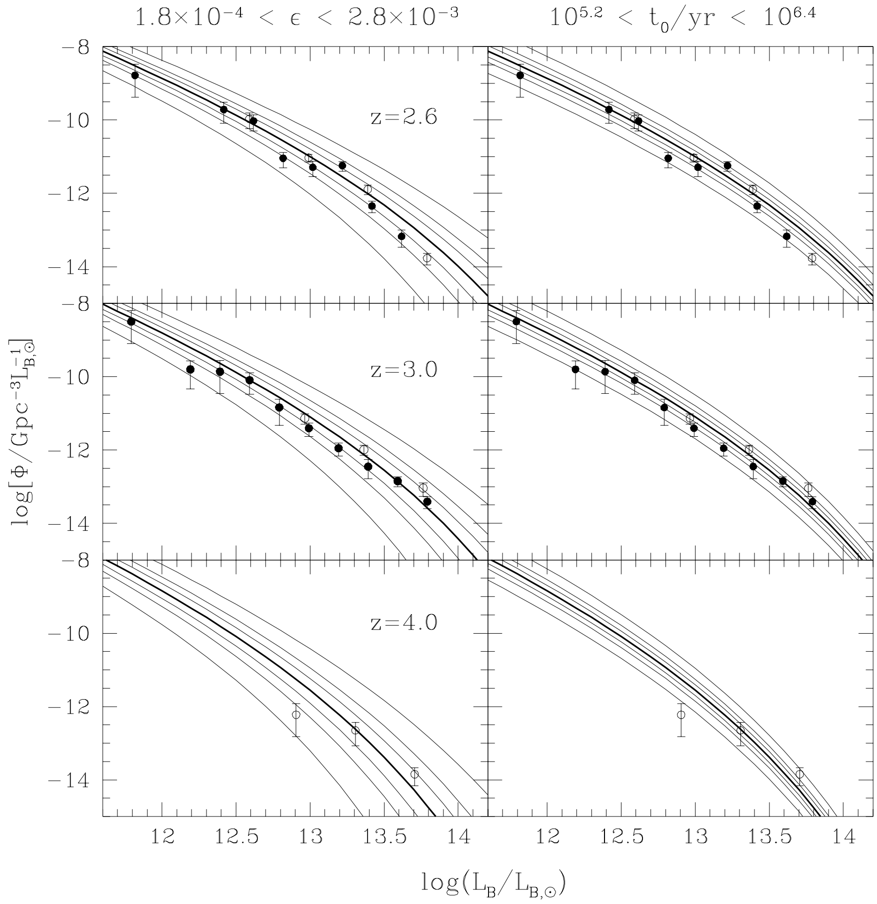


Figure 7.10: *Sensitivity of the derived luminosity function at three different redshifts, $z = 2.6$, 3, and 4, to variations of the two model parameters ϵ (black–hole to halo mass ratio) and t_0 (radiative lifetime of the black hole). The thick lines depict the best–fit model $(\epsilon, t_0) = (10^{-3.2}, 10^{5.8})$. The left column shows the LFs when ϵ is increased by a factor of 1.5, 2, or 4 (upper curves); or decreased by the same factors (lower curves). Similarly, the right column shows the LFs when t_0 is increased (upper curves), or decreased (lower curves) by these factors.*

increase in electron scattering optical depth. The number counts would increase only slightly at the 1nJy threshold, as shown by the dashed lines in Figure 7.11.

Finally, we relax the simplest assumption, and consider the possibility that the black hole formation efficiency is of the nonlinear general form $\epsilon \propto M_{\rm halo}^{\alpha}$ with $\alpha > 0$. This allows black holes to grow in mass as a result of hierarchical merging of halos even after the collapsed fraction approached unity. As an example of a nonlinear efficiency relation, we study the case $\epsilon = \epsilon_0 (M/M_{\rm min})^{0.5}$, where $M_{\rm min} = 10^8 {\rm M}_\odot [(1+z)/10]^{-3/2}$ is the minimum halo mass in which baryons condense. We have found that the quality of the fit to the quasar LF at $2.6 < z < 4$ is no worse than in the linear case with $M_{\rm bh} \propto M_{\rm halo}$. However, in the $\alpha = 0.5$ case, the fit results in a significantly reduced formation efficiency in the shallow potential wells near the $\sim 10^8 {\rm M}_\odot$ threshold, $\epsilon_0 = 10^{-6.2}$, and a longer quasar lifetime, $t_0 = 10^{7.3}$ years. Overall, the extrapolated number of high–redshift faint objects is reduced: with these parameters the reionization redshift is lowered from $z = 12$ to $z = 7.3$ and the high–redshift number counts drop by a factor of 3–20 (see the dotted lines in

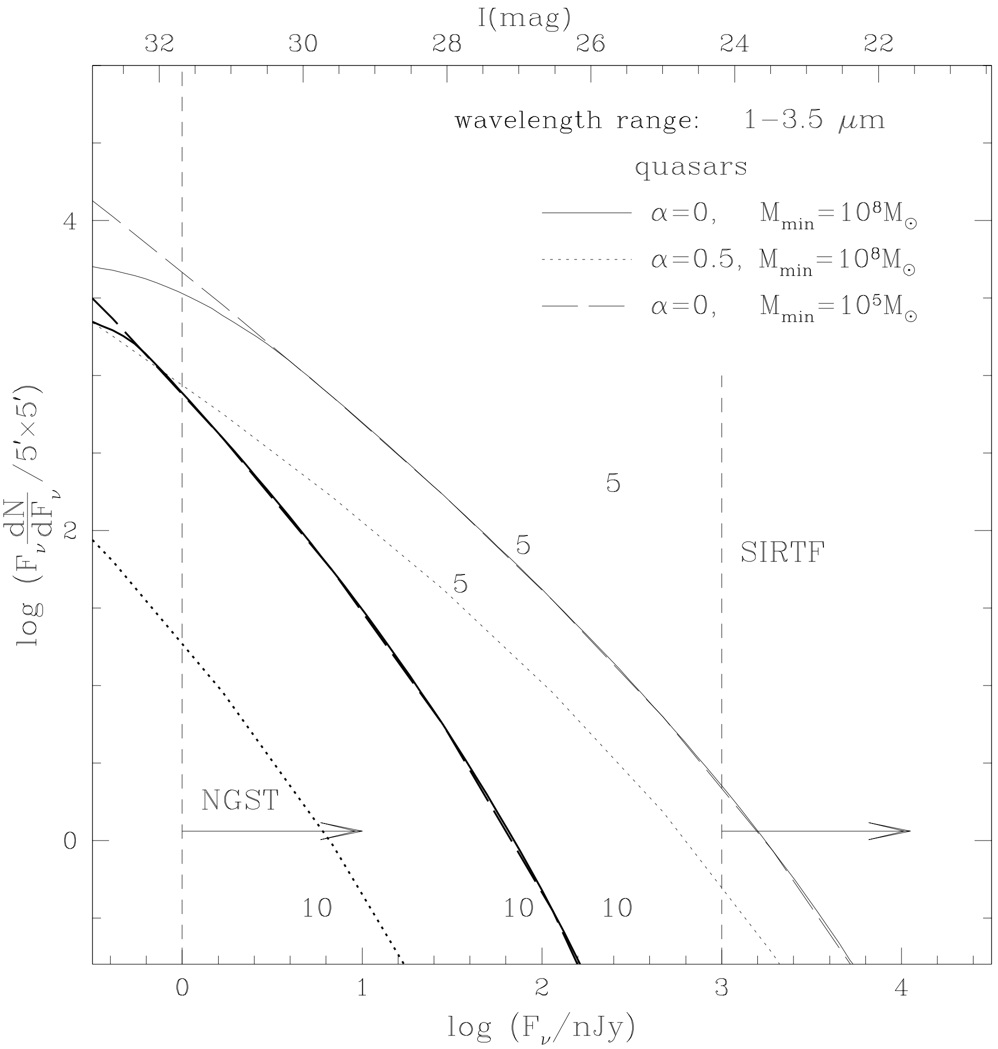


Figure 7.11: *The number per logarithmic flux interval of high–redshift quasars, as in Figure 7.9 (solid lines), for the case where the minimum halo mass is lowered by a factor of* 10^3 *to a value of* $\sim 10^5 \mathrm{M}_\odot$ *(dashed lines), and for the case case where a nonlinear black hole to halo mass relation* $M_{\rm bh} \propto M_{\rm halo}^{1+\alpha}$ *is assumed with* $\alpha = 0.5$ *(dotted lines).*

Figure 7.11).

For a more complete discussion of the dependence of our results in the stellar case on various parameters, including the cosmology, the power–spectrum, $\Omega_{\rm b}$, UV escape fraction, or the stellar IMF, we refer the reader to HL97.

7.7 Conclusions

We extrapolated the quasar luminosity function to faint magnitudes and high redshifts, based on the assumptions that the halo formation rate follows the Press–Schechter theory, and that the quasar light–curve scales linearly with the halo mass, and is otherwise a universal function of time. We demonstrated that a universal light–curve of the form, $L(t) = L_{\rm Edd}\exp(-t/t_0)$, with $t_0 = 6.6 \times 10^5$ yr and $L_{\rm Edd}$ being the Eddington luminosity, provides an excellent fit to observational data for the quasar LF between $2.6 < z < 4.5$ provided that the final black hole mass is a fraction $\sim 5.4 \times 10^{-3}$ of the baryonic mass in each object. This fraction is surprisingly close to the roughly universal ratio between black hole and bulge masses observed in nearby galaxies (Magorrian et al. 1997). Our extrapolation of the quasar LF to high redshifts and low luminosities as presented in Figure 7.4, can be viewed as a natural by–product of this physically motivated fit.

Given the Press–Schechter prediction for the formation rate of halos, and the fact that t_0 is much shorter than the Hubble time, our prediction for the reionization history depends only on the integral of the inferred light–curve, namely on the mean radiative efficiency of the material that made the black hole. The low value inferred for t_0 implies an average radiative efficiency of only $\sim 0.1\%$. This rather low estimate for the efficiency is smaller by two orders of magnitude than the value expected in thin disk accretion; any increase in its value would only enhance the effects we predict. Despite our conservative approach, we predict an extensive population of faint quasars at high redshifts. For illustration, we show in Figure 7.5 the expected evolution in the number density of bright quasars with absolute B magnitude $M_B < -27.5$, along with that of quasars which are two orders of magnitude fainter. Although the density of bright quasars in our models declines rapidly beyond redshift $z \sim 3$ in good agreement with existing observations at $2.6 < z < 4.5$, we predict an abundant population of faint quasars which extends out to redshifts $z \gtrsim 10$, and has not yet been directly observed.

We find that, in general, the effects of early quasars are comparable to those of early stars. However, stellar effects scale with the average IGM metallicity that we used to normalize the star formation efficiency. If the observed values of the C/H ratio in the individual Lyα forest clouds are representative of the universal average C/H ratio, than this ratio implies an IGM metallicity of $\sim$ 0.5–1% solar. However, if the overall carbon abundance is patchy, because it is not well mixed with the rest of the intergalactic gas (as argued by Gnedin 1998), then the true average metallicity may be an order of magnitude lower, 0.1% solar (Songaila 1997). We have found that at the high end of the allowed metallicity range, the effects of stars are stronger than that of quasars, while at the low end of this range, quasars are more effective. In particular, quasars reionize the universe at $z \sim 11.5$, while stars at $9 \lesssim z \lesssim 13$. The Compton y–parameter for the spectral distortion of the CMB is expected to be $1.3 \times 10^{-7} < y_{\rm c} < 1.2 \times 10^{-5}$ from starlight alone, and $4.1 \times 10^{-6} < y_{\rm c} < 2.0 \times 10^{-5}$ if the flux of quasars is added. The direct far-infrared flux from quasars yields a lower limit to the distortion, $y_{\rm c} \gtrsim 3.4 \times 10^{-6}$, even in the absence of any intergalactic dust. The expected number of $z > 10$ quasars per square arcminute, brighter than 1nJy in the wavelength range 1–3.5μm, is $\sim$ 40, while the corresponding number for star clusters is $\sim$ 1–100. Both of these numbers are large enough to show as a distinct population on a typical image by the forthcoming Next Generation Space Telescope.

Chapter 8

Constraints from the Hubble Deep Field on High Redshift Quasar Models

Zoltán Haiman[1], Piero Madau[2], and Abraham Loeb[1]

[Submitted to *The Astrophysical Journal*, May (1998)]

ABSTRACT

High resolution, deep imaging surveys are instrumental in setting constraints on simple semi–analytical structure formation models in Cold Dark Matter (CDM) cosmologies. We show here that the lack of unresolved B–band "dropouts" with $V > 25$ mag in the Hubble Deep Field (HDF) is inconsistent with the expected number of quasars in popular CDM models, in which massive black holes form in halos with a constant universal efficiency. To reconcile the models with the data, a feedback mechanism is needed that suppresses the formation of quasars in halos with circular velocities $v_{\rm circ} \lesssim 75$ km s^{-1}. While the expected number counts could also be reduced: (1) by a steep dependence of the black hole mass on its host–halo mass, $M_{\rm bh} \propto M_{\rm halo}^{1+\alpha}$ with $\alpha \gtrsim 0.5$, or (2) by a shift of the "big blue bump" component to very short wavelengths (~ 200Å) in low–mass quasar spectra, or (3) by using cosmologies in which low–mass halos form at sufficiently high redshifts (such as an open $\Omega_0 = 0.35$ model with a power spectrum index $n = 1.15$), neither of these effects can alone account for the observed lack of detections. If reddening by dust can be neglected at early epochs, consistency with the optical data also requires that the luminous extent of dwarf galaxies at high redshifts be larger than a few percent of the virial radii of their dark matter halos, in order not to predict a large number of additional point–like B–band dropouts. Future deep observations in the J and H bands with NICMOS could reveal several $z \gtrsim 5$ objects per field, and/or provide even stronger constraints on the models than existing B, V, and I data.

[1]Astronomy Department, Harvard University, 60 Garden Street, Cambridge, MA 02138, USA
[2]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

8.1 Introduction

One of the primary goals of models for structure formation in the Universe is to explore the formation process of luminous objects such as quasars or galaxies, and to explain the shape and evolution of their observed luminosity functions (LFs). Recent theoretical work has focused on the formation of galaxies within the framework of hierarchical structure formation in CDM cosmologies, employing either three–dimensional numerical simulations (Gnedin & Ostriker 1997; Zheng et al. 1997), or semi–analytic modeling (Kauffman, Guiderdoni & White 1994; Cole et al. 1994; Baugh et al. 1997; Somerville & Primack 1998; Haiman & Loeb 1997, hereafter HL97). These calculations have been quite successful in reproducing the observed Tully–Fisher relation in the local Universe, the faint galaxy LFs out to moderately high redshifts $z \lesssim 3$, and the recently discovered clustering of Lyman–break galaxies at $z \sim 3$ (Baugh et al. 1997; Governato et al. 1998; Steidel et al. 1998). Hierarchical models were also applied to the more complex problem of quasar formation (Haehnelt & Rees 1993; Small & Blandford 1992), and were shown to reproduce the observed evolution in the quasar LF at redshifts $z \gtrsim 2$ under a minimal set of simplifying assumptions (Haiman & Loeb 1998b, hereafter HL98; see also Haehnelt, Natarajan & Rees 1998).

The above models have not been tested in detail at redshifts much higher than $z \sim 3$ because of the lack of related observational data. Such tests are desirable from the theoretical point of view, since the model predictions are expected to be more robust at higher redshifts when the stars and quasars first condensed out of the well-defined set of initial conditions in the early Universe. The exponentially low abundance of collapsed objects at very early times reduces the probability for complicated merger events. Moreover, the chemical and thermal conditions inside objects which assembled out of the pristine intergalactic gas are better defined before metal enrichment and thermal feedback from multiple generations of stars started to affect the appearance of quasars and galaxies. To date, the deepest optical image of the Universe is in the form of the Hubble Deep Field (HDF, Williams et al. 1996), which surveyed a $2.3' \times 2.3'$ field in four bandpasses down to ~ 30 magnitude. Although this image has revealed a large number of galaxies of different morphological types, a surprisingly large fraction of the sky was seen to remain dark at the HDF detection limit (e.g., Vogeley 1997). The Next Generation Space Telescope (NGST), scheduled for launch in 2007, is planned to have a sensitivity of 1 nJy in the 1–3.5μm range, and should be able to detect globular cluster scale objects out to $z \sim 10$ (see Smith & Koratkar 1998).

The properties of faint extended sources detected in the HDF (Madau et al. 1996) have been well described by detailed semi–analytic galaxy formation models (Baugh et al. 1997). On the other hand, the HDF has revealed only a handful of faint unresolved sources (Méndez et al. 1996), and none with the colors expected for high redshift objects. The purpose of the present paper is to show that the lack of point–like sources in the HDF can be used to constrain models for the formation of cosmic structure at redshifts higher than previously possible ($z \gtrsim 4$). The observational data and color–selection criteria are summarized in § 2 and § 3, and the simplest versions of the semi-analytic CDM models are outlined in § 4. In § 5, we compute the optical and infrared quasar number counts in these models, and show that the predictions from standard scenarios are inconsistent with the observed lack of high–z sources. In § 6, we describe possible modifications of these models that could reduce the predicted counts. In § 7, we discuss the expected galaxy number counts. In § 8, we present our conclusions and argue that future infrared observations in the J and H bands with NICMOS may in principle be used to place even stronger constraints on this type of scenarios.

8.2 High Redshift Quasar Candidates in the HDF

Several surveys of unresolved objects in the HDF have appeared in the literature (Elston, Santiago, & Gilmore 1996; Flynn, Gould, & Bahcall 1996; Méndez et al. 1996). Méndez et al. (1996) were able to identify 14 point–like objects, with only 6 of them at magnitudes fainter than $V = 26.5$[3]. The faintest sources in their sample tend to have rather blue $B-V$ and $V-I$ colors (see also Elston et al. 1996). These authors point out that, while the star–galaxy separation software fails above $V \approx 27.5$, misclassification at fainter magnitudes turns more galaxies into stars than vice versa, as the overwhelming majority of objects in the HDF are galaxies and diffuse and faint features disappear first. Hence the number of point–like sources detected down to $V = 30$ should be considered as an absolute upper limit to the number of stars or quasars observed in this magnitude interval. In analogy with star–forming galaxies, high redshifts ($z \gtrsim 3.5$) quasars are expected to have very red $B-V$ and reddish $V-I$ colors because of continuum absorption and line blanketing from neutral hydrogen along the line of sight. As pointed out by Elston et al. (1996), however, none of the stellar objects in the HDF has the colors expected for high–z quasars. A recent detailed search for faint quasars in the HDF by Conti et al. (1998) confirms the lack of a very red compact population: down to a 50% completeness limit of $V = 29.5$, no $z > 3.5$ quasar candidates were found.

8.3 Identifying Sources at z $\gtrsim$ 3.5

In order to compare our models directly with the observations, we show in Figure 8.1 the location of the nine faint $V > 25$ sources detected by Méndez et al. (1996) on a $B-V$ vs. $V-I$ color-color diagram (empty squares). The solid lines in this figure show the predicted trajectories of model galaxies born at redshifts $3 < z < 4.5$ and passively evolving for 10^9 yrs after an instantaneous burst of star formation. The solid dots show the expected colors of quasars at the same redshifts. Unlike galaxies – which change their colors due to the evolution of their stellar population – we have assumed that quasar aging only changes the overall normalization, but not the shape of their intrinsic spectra (see §4 below for details of our model). While all $z > 3.5$ objects are expected to fall within the trapezoidal region shown in the figure, none of the Méndez et al. (1996) sources appears to have the colors predicted for a QSO or a galaxy at high–z. One caveat to the color–selection method is that some QSOs could have unusual colors. Indeed, the three sources that have $(B-V) > 1$ in the Méndez sample are found close to the border of the QSO selection region. However, all of these three objects are rather bright (V=25, 25.5, 26.3), and if they are indeed QSOs, they are likely to be at low ($z < 3$) redshift.

It is interesting to note, in this context, that is should be possible to identify objects at $z > 4.5$ in much the same way as the $z \sim 4$ galaxies are identified by the B dropout technique. Since at higher redshifts line blanketing and Lyman–continuum absorption are even more pronounced due to the larger number of intervening systems along the line of sight, a straightforward extension of the dropout technique, using I, J, and H data, should work fairly well. In analogy with the $z \gtrsim 3.5$ B dropout galaxies, we expect V, I, and J dropout sources to flag the redshifts $z \gtrsim 5$, 7, and 9. The $B-V$ vs. $V-I$ diagram used to identify $z \sim 4$ star-forming objects

[3]Magnitudes throughout this paper are in the AB system, i.e. $m_{X,AB} = -2.5 \log[f_X/\mathrm{erg\ s^{-1}\ cm^{-2}\ Hz^{-1}}] - 48.6$.

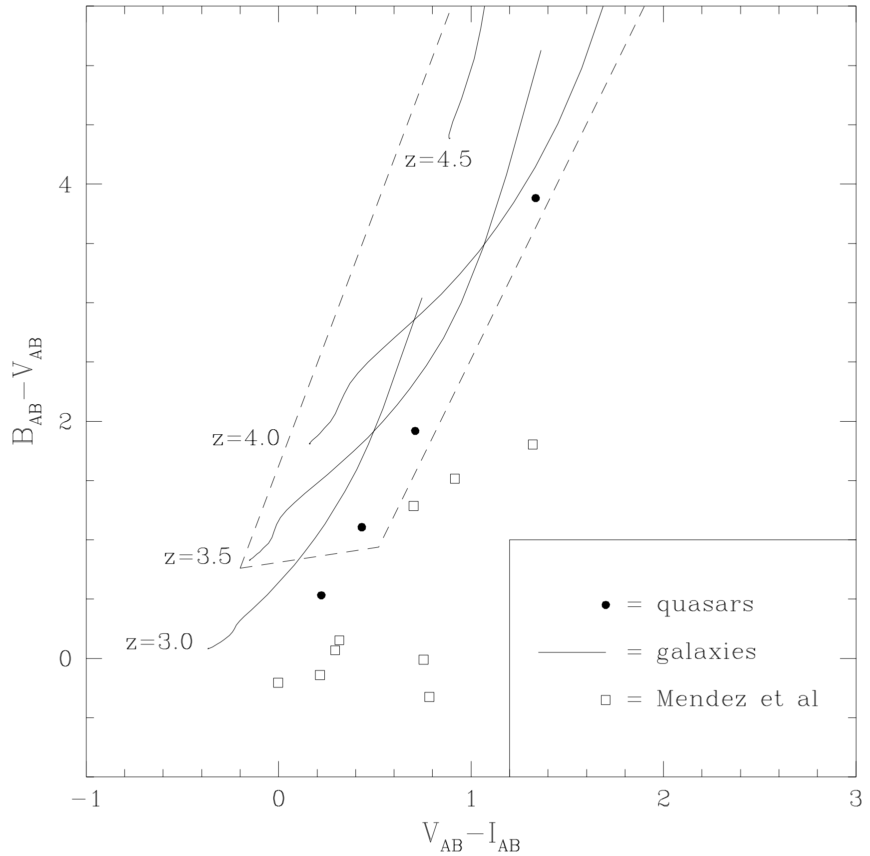


Figure 8.1: *Theoretical $B-V$ vs. $V-I$ color–color diagram for $z>3$ objects. The solid dots show quasars whose intrinsic colors are assumed not to change with cosmic time, while the solid lines show galaxies with ages $0-10^9$ years. The data points (empty squares) are the nine faint ($V>25$), unresolved objects from Méndez et al. (1996). The trapezoidal region enclosed by the dashed lines should selects all unreddened sources with $z\gtrsim 3.5$.*

(Madau et al. 1996) could then be replaced by a $I-J$ vs. $J-H$ color–color diagram as a probe of the redshift range $4.5<z<8$, as illustrated in Figure 8.2. Unreddened galaxies and quasars at $z>4.5$ are expected to be confined to the trapezoidal region indicated in this figure, with negligible contamination from foreground stars.

8.4 Model Description

In modeling the abundance of high–redshift sources, we assume that the formation of dark matter halos follows the Press–Schechter (1974) theory. Hence, the net rate of change in the comoving number density of halos with mass $M_{\rm halo}$ is given by the derivative $\frac{d}{dz}(dn_{\rm c}/dM)$, where $dn_{\rm c}/dM$ is the Press–Schechter mass function (comoving number density per unit mass). This rate includes a negative contribution from the disappearance of small halos in merger events; the net $\frac{d}{dz}(dn_{\rm c}/dM)$ becomes negative when mergers dominate, i.e for masses below a characteristic mass scale $M_*(z)$ at

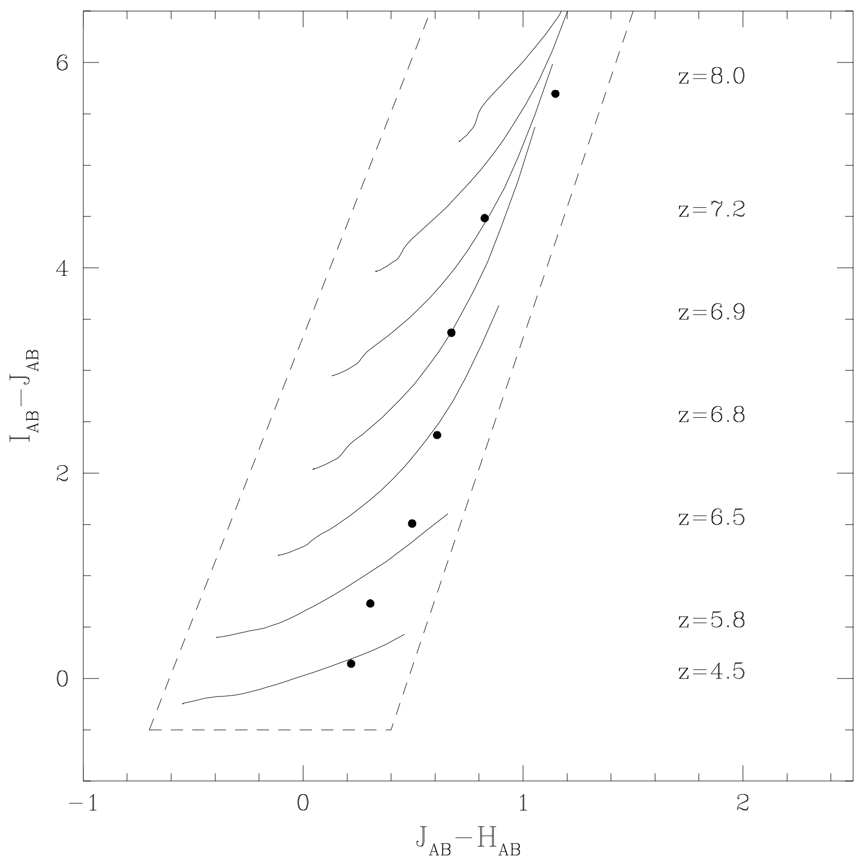


Figure 8.2: *Theoretical* $I-J$ *vs.* $J-H$ *color–color diagram for* $4.5 \lesssim z \lesssim 8$ *objects. The solid dots show the expected colors of quasars, while the solid lines show galaxies with ages* $0-10^9$ *years. The trapezoidal region selects objects with* $z \gtrsim 4.5$.

any given redshift z. Quasars are short–lived and we assume that their formation rate is proportional to the formation rate of new halos. The negative contribution to this rate from merger events should therefore be ignored. For relatively bright quasars at high redshifts (and for all cases in our standard ΛCDM model), the Press–Schechter formation rate is adequate since the typical halo mass $M_{\rm halo}$ is above $M_*(z)$. Otherwise, we make the conservative assumption that no quasars form in halos with a mass $M_{\rm halo} < M_*(z)$ and therefore set $\frac{d}{dz}(dn_{\rm c}/dM) = 0$ (see discussion below).

After each halo collapses and virializes, its gas might condense into stars or a quasar black hole, provided it can cool efficiently from its initial virial temperature. Efficient cooling via collisional excitation of atomic hydrogen requires the halo mass to be at least $M_{\rm min} \sim 10^8 {\rm M}_\odot[(1+z)/11]^{-3/2}$ (HL97). In halos that exceed this mass, a fraction $\epsilon_\star$ of the gas might fragment into stars – forming a dwarf galaxy with a stellar mass of $M_{\rm star} = \epsilon_\star M_{\rm halo}$. Alternatively (or concurrently), a different fraction $\epsilon_{\rm Q}$ of the gas might accrete onto a massive central black hole with a mass $M_{\rm bh} = \epsilon_{\rm Q} M_{\rm halo}$ and exhibit a low–luminosity quasar activity (Eisenstein & Loeb 1995; Loeb 1998). The efficiencies $\epsilon_\star$ and $\epsilon_{\rm Q}$ depend on several parameters, such as the redshift, halo mass, or the initial angular momentum of the gas. However, the simplest assumption to make is that the efficiencies are the same for all halos.

An important effect we need to consider is an external feedback from photoionization, which exerts its influence by heating the gas before it is able to cool and collapse into the dark matter potential wells. Such a feedback is an inevitable consequence of the background UV flux that builds up after the reionization epoch, when the cosmological HII regions have overlapped and the universe is transparent to the ionizing flux from each individual source. Several authors have discussed the consequences of this feedback (Quinn, Katz & Efstathiou 1996; Thoul & Weinberg 1996; Navarro & Steinmetz 1997). According to the spherical simulations of Thoul & Weinberg, a UV background at the level expected from the proximity effect ($J_{21} = 0.1 - 1$; Bajtlik, Duncan & Ostriker 1988) completely suppresses the cooling and collapse of gas inside halos with circular velocities $v_{\rm circ} \lesssim 35$ km s^{-1}, and substantially reduce (by $\approx 50\%$) the mass of cooled baryons in systems with $v_{\rm circ}$ up to 50 km s^{-1}. While this type of feedback has been mainly invoked to suppress the formation of dwarf galaxies (Ikeuchi 1986; Efstathiou 1992; Babul & Rees 1992), it is also natural to expect that a similar mechanism may inhibit the formation, or the fueling, of quasar black holes in shallow potential wells. To mimic these effects, in our standard model we impose a minimum circular velocity $v_{\rm circ} = 50$ km s^{-1} for the halos of both luminous quasars and galaxies that form after the reionization epoch. This constraint results in a minimum halo mass of $M_{\rm min} \sim 10^{10}{\rm M}_\odot[(1+z)/11]^{-3/2}$, two order of magnitudes higher than the cutoff mass obtained from the cooling argument.

The value of $\epsilon_\star$ can be derived from the requirement that the resulting stellar population reproduces the metallicity observed at $z \sim 3$ in the Lyα forest, $10^{-2} \lesssim Z_{\rm Ly\alpha} \lesssim 10^{-3}{\rm Z}_\odot$ (Tytler et al. 1995; Songaila & Cowie 1996; Songaila 1997). Assuming a Scalo (1986) stellar initial mass function (IMF), and using carbon as a measure of metallicity, this yields $0.017 \lesssim \epsilon_\star \lesssim 0.17$ (HL97). Similarly, the value of $\epsilon_{\rm Q}$ for high–redshift quasars can be found by matching the observed LF of quasars at intermediate redshifts, $\phi(L, z)$. Assuming that all quasars shine with a universal light–curve in Eddington units [i.e. that $L(t)$ scales linearly with the black–hole mass, $L(t) = M_{\rm bh}f(t) = \epsilon_{\rm Q}M_{\rm halo}f(t)$], the constant $\epsilon_{\rm Q}$ and the function $f(t)$ can be adjusted so as to produce the observed $\phi(L, z)$ at $2.6 \lesssim z \lesssim 4$ according to

$$\phi(L,z) = \int_z^\infty \int_{M_{\rm min}}^\infty dM_{\rm halo}dz' \frac{d^2n_{\rm halo}}{dM_{\rm halo}dz'}\delta[L - \epsilon_{\rm Q}M_{\rm halo}f(t_{z,z'})], \tag{8.1}$$

where $t_{z,z'}$ is the time elapsed between the redshifts z and z'. To account for the decline of the observed quasar population at redshifts below $z \sim 2.6$ would require additional modeling, but should not effect our conclusions here. The fit to the observational data for $\phi(L, z)$ (in the rest–frame B band) given by Pei (1995) yields $\epsilon_{\rm Q} \sim 6 \times 10^{-4}$ and a light–curve shape that is well approximated by an exponential, $f(t) \sim \exp(-t/t_0)$ with $t_0 = 10^{5.8}$ yr in a flat Λ–CDM cosmology with $\Omega_\Lambda = 0.65$ and a Hubble constant $h = 0.65$. Although the best fit values for the t_0 and $\epsilon_{\rm Q}$ vary with cosmology (see discussion in § 6), the adopted procedure fixes $\epsilon_{\rm Q}$ and t_0 to within a factor of two in any given world model (HL98). We note that t_0 could be increased substantially if black holes formed only in a fraction f of all halos. For example, Eisenstein & Loeb (1995) have argued that only the small subset of initial seed perturbations with sufficiently slow initial rotation can form black holes. In order to recover the fit to the luminosity function, the lifetime would then need to be increased to $\sim t_0/f$.

The Press–Schechter halo formation rate depends on the cosmology and power spectrum. For our "standard" scenario we have adopted the concordance model of Ostriker & Steinhardt (1995), i.e. a flat ΛCDM model with a slightly tilted power spectrum $(\Omega_0, \Omega_\Lambda, \Omega_{\rm b}, h, \sigma_{8h^{-1}}, n)$=(0.35, 0.65, 0.04, 0.65, 0.87, 0.96). Convenient expressions for the differential volume element, luminosity distance, and time–redshift relation in this model were given in terms of elliptic integrals by Eisenstein (1997); an accurate fitting formula for the growth function is given by Carroll et al. (1992). We also consider two open models without a cosmological constant, one with Ω_0=0.35 and the other with $\Omega_0 = 0.2$. In order to remain consistent with the power–spectrum normalization on both COBE and cluster scales (Viana & Liddle 1996), we change the spectral index to n=1.15 and 1.2, respectively, in these models.

For our quasar template spectrum, we have used the average spectrum derived from the observations of 47 quasars by Elvis et al. (1994). The stellar template spectra have been computed from the low–metallicity ($Z = 4 \times 10^{-4}$) population synthesis models of Bruzual & Charlot (1996), assuming an instantaneous burst of star formation and a Scalo (1986) IMF. Note that a Salpeter IMF produces about twice as many UV photons as the Scalo IMF, for the same level of carbon enrichment of the intergalactic medium (IGM). We find that adopting a Salpeter IMF would increase our predicted number counts down to $V \sim 30$. In our standard model, the increase is negligible at $V = 30$, while it is an order of magnitude at $V = 22$.

8.5 Number Counts

In both the quasar and galaxy models, the number of sources located between redshifts $z_{\rm min}$ and $z_{\rm max}$ per unit solid angle, with observed visible magnitude between V and $V + dV$, is given by

$$\frac{dN}{d\Omega dV}(V, z_{\rm min}, z_{\rm max}) = \int_{z_{\rm min}}^{z_{\rm max}} dz \left(\frac{dV_{\rm c}}{dz d\Omega}\right) n_{\rm c}(z, V), \tag{8.2}$$

where $dV_{\rm c}/dzd\Omega$ is the comoving volume element per unit redshift per unit solid angle (Eisenstein 1997), and $n_{\rm c}(z, V)$ is the comoving number density of objects at a redshift z, whose magnitude is observed to be between V and $V + dV$.

Based on the halo formation rate $d^2n/dMdz$, $n_{\rm c}$ is given by a sum over halos of different ages that exist at each redshift,

$$n_{\rm c}(z, V) = \int_z^\infty dz' \frac{dM_{\rm halo}}{dV}(z, z', V) \left.\frac{d^2 n_{\rm c}}{dM dz'}\right|_{M_{\rm halo}(z,z',V),z'}, \tag{8.3}$$

where the factor $dM_{\rm halo}/dV$ converts the number density per unit mass interval to number density per unit flux interval. The V magnitude in the AB system is given by

$$V = -2.5 \log f_V - 48.60, \tag{8.4}$$

where the observed V–band flux $f_V(z, z', M_{\rm halo})$ from a halo at redshift z with a mass $M_{\rm halo}$ that was formed at $z' \geq z$ equals (in units of erg sec^{-1} Hz^{-1} cm^{-2}):

$$f_V(z, z', M_{\rm halo}) = \frac{(1+z)\epsilon M_{\rm halo}}{4\pi d_{\rm L}(z)^2} \left[\int_0^\infty d\nu T_V(\nu) \exp\left[-\tau(\nu, z)\right] j(\nu(1+z), t_{z,z'})\right] \times \left[\int_0^\infty d\nu T_V(\nu)\right]^{-1}. \tag{8.5}$$

Here $d_{\rm L}(z)$ is the luminosity distance; $\epsilon = \epsilon_{\rm Q} \equiv 6 \times 10^{-4}$ or $\epsilon = \epsilon_\star \equiv (0.017 - 0.17)\Omega_{\rm b}/\Omega_0 = 0.002 - 0.02$ is the fraction of halo mass in the form of the central black hole or stars, respectively; $\tau(\nu, z)$ is the effective optical depth associated with intervening Lyα forest clouds and Lyman–limit systems between redshifts

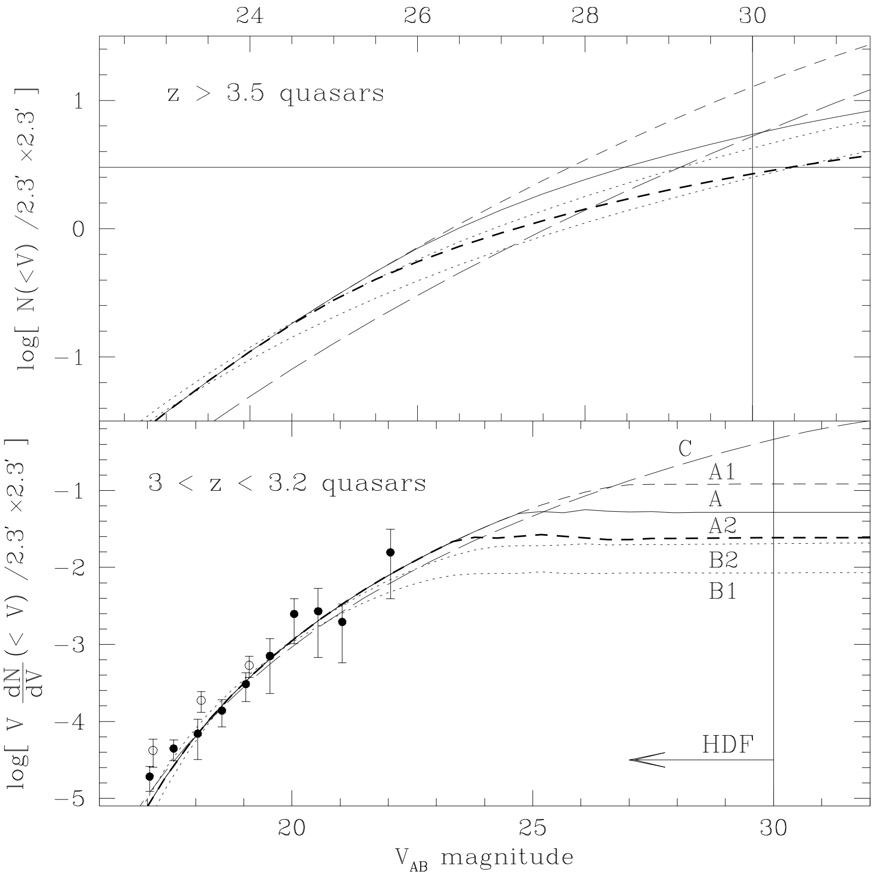


Figure 8.3: *V-band number counts for high–redshift quasars. The lower panel shows differential counts for* $3 < z < 3.2$*, as well as data adapted from Pei (1995). The upper panel shows cumulative number counts for* $z > 3.5$*. The curves labeled A show predictions of our standard ΛCDM model with* $\mathrm{M_{bh}} \propto \mathrm{M_{halo}}$*. The curves labeled A1 show the counts in a model with no UV feedback. The curves labeled A2 correspond to a model with minimum halo circular velocities* $v_{\rm circ} \geq 75$ km s^{-1}*. The curves labeled B1 and B2 correspond to open CDM models with* Ω_0*=0.35 and 0.2, respectively, and the curves labeled C corresponds to a "tilted" model with* $\mathrm{M_{bh}} \propto \mathrm{M_{halo}^{1.5}}$.

0 and z at the observed frequency ν (Madau 1995); $T_V(\nu)$ is the frequency–dependent transmission of the V_{606} filter[4]; $j(\nu, t_{z,z'})$ is the template luminosity per unit (stellar or black hole) mass for either stars or quasars, in units of erg sec^{-1} Hz^{-1} M$_\odot^{-1}$; and $t_{z,z'}$ is the time elapsed between redshifts z' and z. The time dependence of $j(\nu)$ is obtained from our fitted light–curve function $L(t)$ for quasars, and from tabulated evolutionary tracks for stars (Schaller et al. 1992; see HL97 for details). We suppressed the emergent flux from all objects at emission frequencies above Lyα before reionization occurs (assuming sudden reionization at the redshift $z = 13$ or 9 for stars, and $z = 11.5$ for quasars, see HL98), since the Gunn–Peterson optical depth of HI at these frequencies is exceedingly high prior to reionization.

The expected counts of $z > 3.5$ quasars in the V band are shown by the solid lines, labeled A, in Figure 8.3. The bottom panel also shows the observational data

[4]http://www.stsci.edu/ftp/instrument_news/WFPC2/Wfpc2_faq/wfpc2_phot_faq.html.

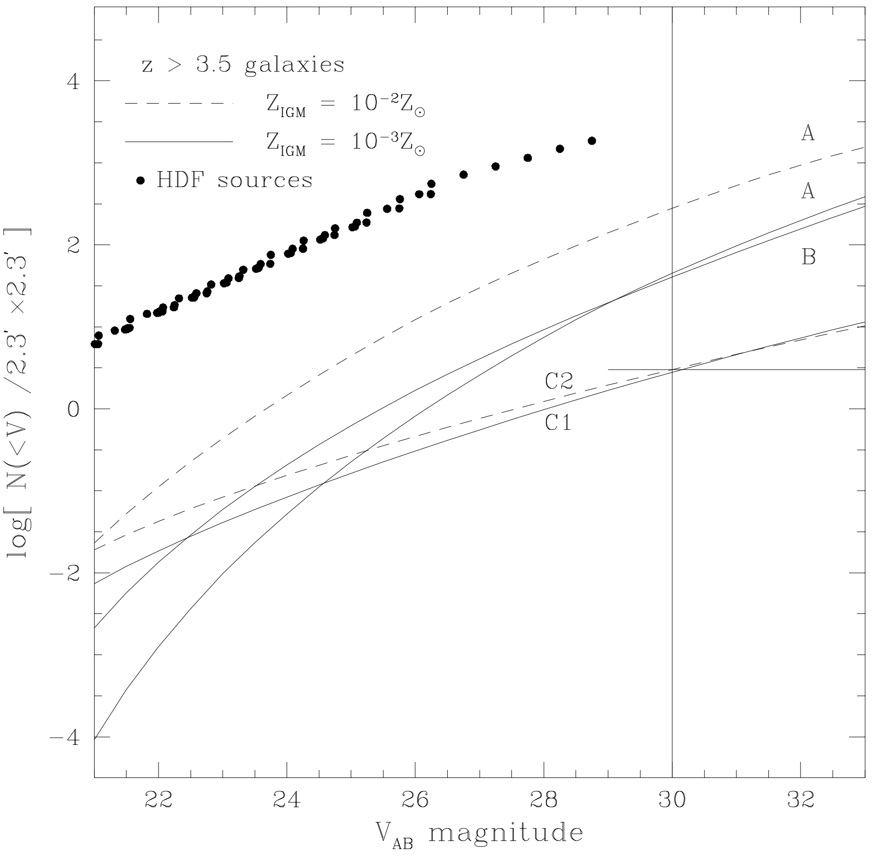


Figure 8.4: *V counts for high–redshift galaxies in the simplest model with* $M_{\rm star} \propto M_{\rm halo}$. *The solid lines correspond to low star formation efficiencies* ($M_{\rm star} = 0.017 M_{\rm halo}$) *that produce an IGM metallicity of* $10^{-3} Z_\odot$, *and the dashed lines correspond to a star formation efficiency which is 10 times higher. The curves labeled A correspond to our standard model. The curve labeled B shows the counts in the open CDM model. The curves labeled C1 and C2 correspond to models with* $\mathrm{M_{bh}} \propto \mathrm{M}^3_{\rm halo}$ *(for* $Z_{\rm Ly\alpha} = 10^{-3}\mathrm{Z}_\odot$*) and* $\mathrm{M_{bh}} \propto \mathrm{M}^{3.5}_{\rm halo}$ *(for* $Z_{\rm Ly\alpha} = 10^{-2}\mathrm{Z}_\odot$*), respectively. The solid dots show all extended sources detected in the HDF in the V band (Williams et al. 1996).*

points from the B band (Pei 1995), converted to V counts using the shape of our template spectrum. By construction, the differential model number counts for quasars in the narrow redshift interval $3 < z < 3.2$ are consistent with the data between $17 < V < 22.5$. The upper panel in this figure shows the cumulative counts of $z > 3.5$ quasars, and reveals that down to the $V \sim 30$ limit of the HDF, our standard model predicts about 6 objects. Since no unresolved sources at $z > 3.5$ have been found in the HDF, a model which is consistent with the HDF data at the 5% level should predict no more than 3 such high-z QSOs. Our standard quasar scenario therefore appears to overpredict the number of faint point–like sources by a factor of ~ 2. Although the existing data on the LF only constrains the individual quasar model parameters, such as $\epsilon_{\rm Q}$ and t_0, to a factor of ~ 2, the $z > 3.5$ number counts are predicted to a much better accuracy, since the model is calibrated by simultaneously adjusting $\epsilon_{\rm Q}$ and t_0 to reproduce the quasar LF at $z \sim 3$. As a result, the overprediction of the quasar counts by a factor of 2 is significant. It is important to remember here that in our standard model, we have eliminated quasars in halos with

$v_{\rm circ} < 50$ km s^{-1} to mimic the effect of photoionization heating. The thin dashed curves in Figure 8.3, labeled A1, show the number counts in a model without this UV feedback, and reveal a prediction of ~ 15 objects to $V = 30$. The overprediction in this model would be rather severe, a factor of ~ 5.

The expected number counts of $z > 3.5$ galaxies in our standard model are shown by the curves labeled A in Figure 8.4. The solid and dashed lines bracket the expected metallicity range of the IGM at $z \approx 3$, $10^{-3}{\rm Z}_\odot < {\rm Z} < 10^{-2}{\rm Z}_\odot$. Depending on the metallicity calibration, the total number of galaxies observable down to $V \sim 30$ mag is between 40 and 280. This exceeds the predicted quasar counts by a factor of $7-50$, and is $10-100$ larger than what would be allowed for point–sources in the HDF. However, depending on their assumed intrinsic sizes, some or most of the $z > 3.5$ galaxies could actually be resolved (see discussion below). We also note that recent revisions of the C/H ratio in these absorbers prefer lower metallicities, closer to $10^{-3}{\rm Z}_\odot$ (Songaila 1997).

8.6 Reducing the z > 3.5 Quasar Counts

As the simplest scenarios appear to overproduce the faint V counts of quasars, the question naturally arises: how should one modify the models to reduce the expected number of faint objects? In general, the theoretical number counts are sensitive to the underlying cosmology and power spectrum, the relation between the halo mass and the emitted luminosity, as well as the time–evolution of each object, including its merger history.

The simplest explanation for the lack of faint quasars would be a more effective UV feedback mechanism, resulting in a higher minimum halo size for luminous objects[5]. In Figure 8.3, the lines labeled A2 show the number counts of quasars in a model with $v_{\rm circ} > 75$ km s^{-1}. The top panel of this figure demonstrates that with this cutoff value of $v_{\rm circ}$, the $z > 3.5$ counts are reduced to 3 objects, i.e. consistent with the observed lack of detections. This required cutoff value is higher than the values obtained from spherical simulations (Thoul & Weinberg 1996). With a UV background at the level of $0.3 < J_{21} < 1$, the the cooled baryon fraction is fully suppressed for halos with $v_{\rm circ} \lesssim 35$ km s^{-1}, suppressed by 50% for $v_{\rm circ} \lesssim 50$ km s^{-1}, and unaffected for circular velocities above $v_{\rm circ} \gtrsim 75$ km s^{-1}. However, using 3–D simulations, Navarro & Steinmetz (1997) find the UV feedback to be more effective, and their circular velocity thresholds are $v_{\rm circ} = 80 - 200$ km s^{-1}, exceeding the required value $v_{\rm circ} = 75$ km s^{-1} derived here.

A second straightforward approach to reduce the faint counts is to make the slope of the luminosity function shallower. When the. luminosity of each object is steady and proportional to the halo mass, $L \propto M$, the LF is related to the mass function simply through $\phi(L) = dn_{\rm c}/dM \times dM/dL \propto dn_{\rm c}/dM$, i.e. the LF has the same slope as the Press–Schechter mass function. At low masses it is $dn_{\rm c}/dM \propto M^{-2+(n-1)/6}$, independent of cosmology and only very weakly dependent on the power spectrum index n. By introducing a non–linear scaling between luminosity and halo mass, $L \propto M^{1+\alpha}$ (see discussion below) one could tune the LF to have a shallower slope, $\phi(L) \propto L^{-1-1/(1+\alpha)}$.

[5]Alternatively, a minimum black hole mass could be caused by an internal feedback when the quasar luminosity exceeds the rate that unbinds the gaseous disk on a dynamical time (Haehnelt, Natarajan & Rees 1998).

In reality, each source shines only for a limited amount of time Δt, and this can be much shorter than the Hubble time. This results in $\phi(L) \sim \Delta t \times dn_c/dMdt \times dM/dL(t)$, where $dM/dL \propto \exp(-t/t_0)$ for our exponential lightcurve. Here, a complication arises because at masses below M_*, the time derivative of the mass function is negative, $dn_c/dMdt < 0$. Physically, this derivative is the sum of a positive term (the formation rate of new halos with mass M), and a negative term (the disappearance rate of existing halos due to mergers), and while the formation rate exceeds the merger rate for $M > M_*$, the latter dominates for $M < M_*$. The exact ratio between the formation and merger rates requires a scheme that goes beyond the Press–Schechter theory and describes the merging history of halos (Lacey & Cole 1993; Kauffmann & White 1993). For sake of simplicity, we will assume here that no halos form below the threshold mass M_*, and that no halos are destroyed by mergers above this mass, i.e. we simply set the derivative $dn_c/dMdt$ to zero for $M < M_*$ and to the Press-Schechter value for $M > M_*$. Although this assumption is ad–hoc, it can only underestimate the number of objects, both above and below M_*. Therefore, the reduction in the number counts we obtain here should be considered as the minimum amount of change that is necessary to fit the observations.

Under the above prescription, the LF flattens to $\phi(L) \propto 1/L$ at luminosities corresponding to $M < M_*$. The cosmological parameters $(\Omega_0, \Lambda, \mathrm{H}_0)$ and the power spectrum parameters $(\sigma_{8\mathrm{h}^{-1}}, n)$, flatten the number counts at $V \sim 30$ (VdN/dV=const) in models where low–mass halos form at sufficiently high redshift. Successful models raise the threshold mass M_* at $3 \lesssim z \lesssim 6$ to high enough values that $V \sim 30$ objects become associated with sub–M_* halos. For example, at $z \sim 3$, the halo mass corresponding to $V \sim 30$ in our models is $\sim 10^{8-9} M_\odot$. In general, M_* is high in models with large values of $\sigma_{8\mathrm{h}^{-1}}$, n, and Ω_0, and low values of Λ. These parameters are related by the normalization of the power spectrum on the COBE (X–ray cluster) scales, which requires $\sigma_{8\mathrm{h}^{-1}}$ to decrease (increase) with decreasing Ω_0. In Figure 8.3, we show two $\Lambda = 0$ open models, in which $\Omega_0 = 0.35$, and 0.2 (lines labeled B1 and B2, respectively). The values of ($\sigma_{8\mathrm{h}^{-1}}$ and n) in these models which are consistent with both the COBE and cluster normalizations are (0.87, 1.15) and (0.87, 1.2), respectively. Note that the values of t_0 and ϵ_Q also need to be readjusted in these cosmologies. We find (t_0, ϵ_Q)=($10^{6.2}$yr, $10^{-3.7}$) for $\Omega_0 = 0.35$ and (t_0, ϵ_Q)=($10^{6.8}$yr, $10^{-3.9}$) for $\Omega_0 = 0.2$. The lines in the bottom panel labeled B1 and B2 show that the $z \sim 3$ number counts indeed flatten at $V \sim 23$ and $V \sim 24$ in these models. The upper panel of this figure also shows that in the $\Omega_0 = 0.35$ model, the cumulative $z > 3.5$ counts at $V \sim 30$ are lowered to the required level of ~ 3 objects. While the $z > 3.5$ counts in the $\Omega_0 = 0.2$ model are also reduced, the predicted number of quasars at the HDF limit is still too large (~ 5) in this case.

A third alternative approach for reducing the faint $z > 3.5$ counts is to make the LF decline faster at high redshifts. While the quasar LF is reasonably well determined at $z \sim 3$, it is significantly more uncertain around $z \sim 4$ (see Figs. 1 and 2 in Pei 1995). One could exploit this uncertainty and modify the models so that the predicted number of $z \sim 4$ objects is at the low end of the allowed range. We have found that the redshift evolution can not be made sufficiently steep by changing the cosmology or the power spectrum, while still maintaining a good fit to the LF around $z \sim 3$. However, the evolution can be made steeper if the halo mass associated with a fixed luminosity is increased, since the formation rate of higher mass halos evolves more rapidly. To achieve this change, we introduce a simple dependence of luminosity on halo mass, $L = \epsilon_0 (M_{\mathrm{halo}}/10^8 \mathrm{M}_\odot)^{1+\alpha}$ with $\alpha > 0$. A similar scaling, including an additional redshift dependence, $L \propto M_{\mathrm{bh}} \propto v_{\mathrm{vir}}^5 \propto M_{\mathrm{halo}}^{5/3}(1+z)^{5/2}$ has been recently considered by Haehnelt, Natarajan & Rees (1998). Our adopted scaling lacks the factor $(1+z)^{5/2}$, and therefore predicts a steeper redshift evolution and a slightly smaller number of $z > 3.5$ objects. Since most objects at $V \sim 30$ are located at $z < 6$, the difference in the number counts predicted by the two models in the relevant magnitude range is small ($\lesssim 15\%$).

For any given value of α, the fitting procedures described in § 4 can be repeated to obtain the values of ϵ_0 and t_0. Although by definition, our procedure for quasars fits the LF at $2.6 < z < 4$, increasing α means that the characteristic mass in the LF is increased, and the total number of predicted faint $z > 3.5$ quasars is reduced. We have experimented with values of $0 < \alpha < 1$, and found that for $\alpha \gtrsim 0.5$, the fitting procedure for quasars fails, due to the implausibly steep redshift evolution of the model LF already for the observed redshift interval $2.6 < z < 4$. Note that much stronger deviations from the linear relation $M_{\rm bh} \propto M_{\rm halo}$ would also be mildly inconsistent with observational data for local galaxies (Magorrian et al. 1997; McLeod 1997). The largest plausible reduction in the faint counts can therefore be achieved using $\alpha = 0.5$. In this case, we obtain $M_{\rm bh}/M_{\rm halo} = 8 \times 10^{-7}(M_{\rm halo}/10^8 {\rm M}_\odot)^{0.5}$ and $f(t) = \exp(-t/3\times10^6{\rm yr})$, and predict the number counts shown by the curves labeled C in Figure 8.3. The bottom panel shows that the quality of the fit to the $z \sim 3$ LF at $17 < V < 22.5$ is comparable to the fit in our standard model. The top panel shows that relative to model A, the $z > 3.5$ counts decrease by a factor of 2.5, to ~ 6 objects. However, even with the largest allowed value of the exponent, the $z > 3.5$ counts remain a factor of ~ 2 too high.

A fourth effect that could contribute to the deficit of observed faint quasars is the expected shift in the "big blue bump" (BBB) component of the quasar spectrum to higher energies for the lower mass black holes which populate the Universe at high redshifts. In thin accretion disk models (e.g., Peterson 1997) the emission temperature, and hence the peak wavelength of the BBB, scales with the black hole mass as $\lambda_{\rm peak} \propto M_{\rm bh}^{0.25}$. In our present template spectrum, which is based on data from local ($z \lesssim 1$) quasars (Elvis et al. 1994), the BBB peak is tentatively located at 1000Å (see HL98). Since the characteristic black hole might be smaller by several orders of magnitude at $z \sim$ 4–6 than at $z \lesssim 1$, the BBB feature might shift by a factor of a few in wavelength. Such a shift would reduce the observed V band flux by a factor of ~ 7, and the rest–frame B band flux (which we use to calibrate our model) by a factor of ~ 2. In particular, if the BBB in high–redshift quasars is blueshifted to 2–300Å, then the faint V quasar counts would be reduced by a factor of $\sim$2, not quite sufficient to bring the number counts into agreement with the HDF data.

8.7 Number Counts of $z > 3.5$ Galaxies

The predicted number of $z > 3.5$ galaxies in our models at the HDF limits is a factor of 7–50 higher than the number of quasars. However, faint galaxies could have sufficiently large angular sizes to appear extended even at cosmological distances, in which case the overprediction relative to the number of point sources would not be relevant.

A simple estimate shows that this might indeed be the case: to estimate what fraction of the $z > 3.5$ galaxies may have angular diameters above the $0.1''$ pixel size of the three wide field cameras (WFC), we assume that each collapsed halo has a $1/r^2$ density profile, with an overdensity of 50 at the edge of the halo relative to the average background density, and that the luminous region of the galaxy extends out to 5% of the virial radius of the halo (see Navarro, Frenk & White 1995).

With these assumptions, we find that for the low star formation efficiency (producing an average metallicity of $Z_{\mathrm{Ly}\alpha} = 10^{-3}\mathrm{Z}_\odot$ in the intergalactic medium), only 0.5% of the $z > 3.5$ objects would be unresolved at $V = 29.8$. As a result, the HDF is expected to contain no point–like galaxies to this magnitude. The unresolved fractions around the HDF sensitivity limit are significantly higher for a high star formation efficiency ($Z_{\mathrm{Ly}\alpha} = 10^{-2}\mathrm{Z}_\odot$). In this case, a fraction equal to 2%, 9%, 18%, and 24% of the $z > 3.5$ objects would be unresolved at $V = 27.5$, 28.0, 28.9 and 29.8, respectively. Therefore, given the expected intrinsic sizes of the galaxies, the overprediction problem disappears at low star formation efficiencies, but is not fully resolved if the star formation efficiency is higher and the apparent angular size of galaxies smaller. This could be interpreted as support for a low star formation efficiency, and a corresponding low average metallicity of the IGM at $z \sim 3$.

Supposing that the stars occupy a region much smaller than 5% of the virial radius of each halo, and that most $z > 3.5$ galaxies are therefore unresolved, it is interesting to examine the sensitivity of the number counts to the underlying cosmology and to the relation between M_{star} and M_{halo}. In general, we have found that it would be more difficult to reconcile the number counts of galaxies with the HDF data than that of quasars. There are two reasons for this result. First, the number counts of galaxies are higher to begin with. Second, the model galaxies are ~ 2 orders of magnitude less luminous than quasars, so that the same V magnitude corresponds to a ~ 100 larger halo mass. The differential number counts flatten at the fixed mass M_*; for galaxies this mass corresponds to $V \sim 30$ at $z = 4.5$.

As an example, the curve labeled B in Figure 8.4 shows the galaxy counts in the $\Omega_0 = 0.35$ open CDM model. The counts are somewhat less steep than in the Λ model (the solid curve labeled A). Although the flattening of the mass function at $M < M_*$ reduces the faint counts, the effect is negligible small at $V \gtrsim 30$. The figure also shows the results of changing the dependence of the star formation efficiency on the halo mass (changing α). When the star formation efficiency is normalized to $Z_{\mathrm{Ly}\alpha} = 10^{-3}\mathrm{Z}_\odot$, then $\alpha = 2$ and $\epsilon_0 = 3.3 \times 10^{-9}$ are required for consistency with no detections (see the line labeled C2). Similarly, when $Z_{\mathrm{Ly}\alpha} = 10^{-2}\mathrm{Z}_\odot$ is assumed, $\alpha = 2.5$ and $\epsilon_0 = 7.3 \times 10^{-12}$ are required (C1). As the figure shows, the slopes of the latter number count curves are flattened considerably. For reference, we show also the number of counts of all extended sources detected in the HDF in the V band. These counts have a shallow slope and demonstrate that at the $V = 30$ limit only a small fraction of all detected objects are expected to be at $z > 3.5$. Finally, we note that an additional reduction of the galaxy counts could also be caused by dust extinction. Estimates based on the Calzetti et al. (1994) extinction curves indicate that the number of high redshift ($z \lesssim 3$) sources could be as much as ~ 7 times higher than currently detected (Pettini et al. 1997), when dust reddening is taken into account. Given all of the above uncertainties, we conclude that, unlike in the case of quasars, there is no significant discrepancy between the expected galaxy counts and the lack of faint point sources. This is true as long as the stars occupy a region extending out to a few percent of the virial radius of their host halo, and the star–formation efficiency is close to the value expected from a low average Lyα forest metallicity, $Z_{\mathrm{Ly}\alpha} = 10^{-3}\mathrm{Z}_\odot$.

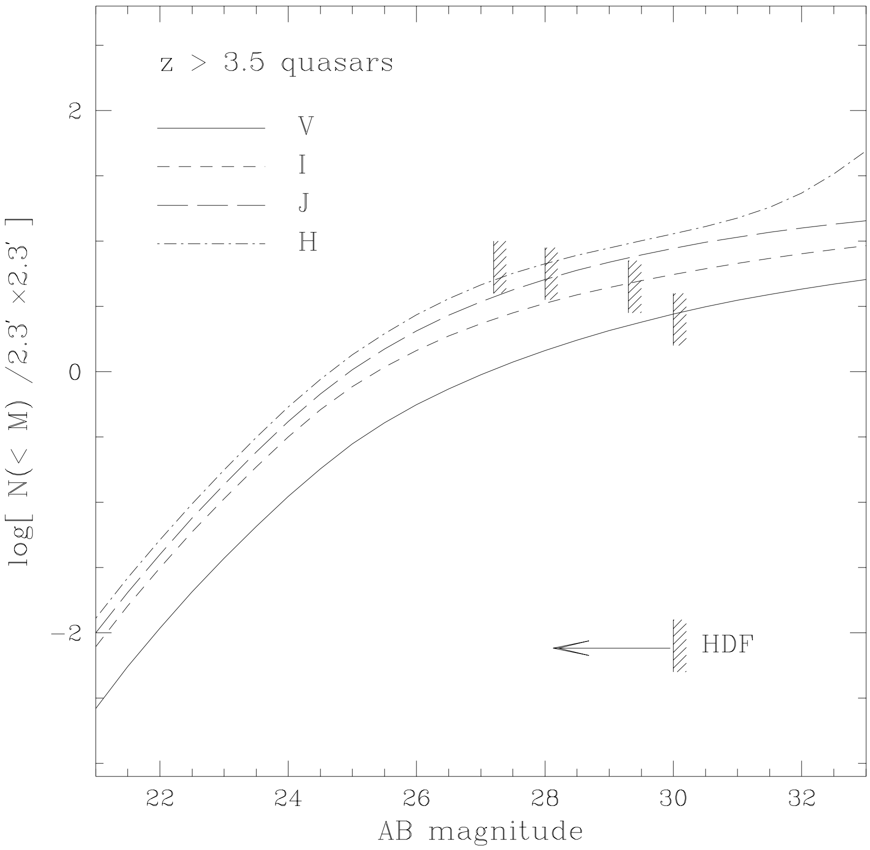


Figure 8.5: *V, I, J, and H counts for quasars in the Λ CDM model with Ω_0=0.35. This model is consistent with HDF data in the V band at the 5% level. Sensitivities shown for each band are for a fixed signal–to–noise ratio and exposure time.*

8.8 Conclusions

The observed lack of faint, unresolved, and red sources in the HDF down to $V \approx 30$ requires models to predict no more than ~ 3 point–like high-z objects to this limiting magnitude. This is in conflict with the predicted number ($\gtrsim 15$) of high redshift quasars in our simplest hierarchical model, which assumes a universal black hole formation efficiency in a ΛCDM cosmology. In this cosmology, the discrepancy can be resolved if black holes do not form, or are not fueled, in halos with circular velocities $v_{\rm circ} \sim 75$ km s^{-1}. An alternative partial solution can be obtained in cosmologies with early structure formation (so that $M_* \gtrsim 10^{11}\mathrm{M}_\odot$ by $z = 3$), but only by postulating that new black holes do not form in halos below the nonlinear mass–scale, $M_*(z)$. An example for such a model is the $\Omega_0 = 0.35$ open CDM model, with a power spectrum slope $n = 1.15$ and normalization $\sigma_{8h^{-1}} = 0.87$. The $z > 3.5$ number counts can be further reduced by introducing a black hole formation efficiency $M_{\rm bh}/M_{\rm halo}$ that depends on the halo mass as $\propto M_{\rm halo}^{0.5}$. Similarly, a shift of the big blue bump component to shorter wavelengths in the spectra of faint quasars could also help reduce the predicted counts. However, neither of these effects could explain, by itself, the observed lack of detections.

The models predict $\sim$ 3–20 times more $z > 3.5$ galaxies than quasars at

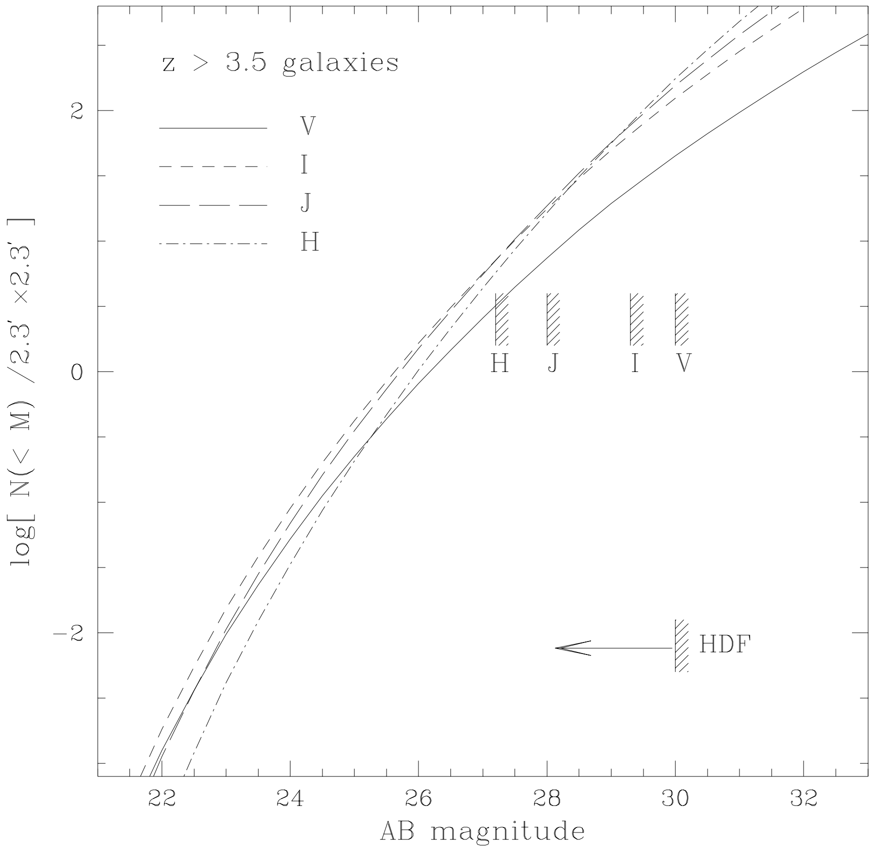


Figure 8.6: *Theoretical V, I, J, and H counts for $z > 3.5$ galaxies in our standard ΛCDM model with $Z_{\rm Ly\alpha} = 10^{-3}Z_\odot$. Since most galaxies at $V < 30$ mag may be resolved, this model could still be consistent with the HDF images.*

$V = 30$. If the star formation rates in these galaxies are adjusted to produce an average Lyα forest metallicity of $10^{-2}Z_\odot$ by $z = 3$, then more than 10% of all galaxies could appear unresolved at the 0.1″ pixel size of the WFC, possibly leading to another discrepancy with the observed lack of detections. The situation changes if the star formation rate is lowered by an order of magnitude (normalized to produce a $10^{-3}Z_\odot$ average IGM metallicity at $z \sim 3$), since then almost all of the predicted $z > 3.5$ galaxies are expected to be extended, rather than point sources.

Future observations at longer infrared wavelengths, such as the J and H infrared bands, would be better suited to the study of the $z \gtrsim 4.5$ universe, as they allow bigger volumes to be probed. In Figure 8.5, we show the expected number counts of quasars in the V, I, J, and H bands for the model which is consistent with the optical HDF data. An increasingly larger number of objects are predicted to be detected in the I, J, and H bands as these are affected by intervening absorption only at extreme redshifts. These number counts could be somewhat reduced since faint, compact (compared to the point spread function) WFC sources lose signal–to–noise when observed by NICMOS, due to the poorer angular resolution of the infrared images. Similar to Figure 8.5, Figure 8.6 shows the expected number of galaxies in the infrared bands. At 30 mag, the gain in observing galaxies redward of the I-band is not as significant as for quasars, because the intrinsic spectra of unreddened galaxies, unlike those of quasars, decreases as a function of wavelength at $z \sim 4$ (although the merit of infrared bands relative to V would be increased at still higher redshifts). The redshifts of either type of sources can be obtained using a straightforward analogy to the B dropout technique, applied to the $I - J$ vs. $J - H$ color–color diagram shown in Figure 8.2. Future deep observations with NICMOS in these bands of a field comparable in size to the HDF would be able to either reveal tens of ultra–high redshift quasars, or else place tighter constraints on quasar models than currently possible using V and I data.

Chapter 9

Summary

According to our present understanding, the universe started from a nearly uniform initial state, with only very small inhomogeneities. Standard cosmological models quantify the subsequent evolution of these inhomogeneities due to gravity, and provide a general framework for the formation of structure on a wide range of scales. Observations of the cosmic microwave background (CMB) radiation and the largest scale distribution of galaxies have singled out a class of models, called cold dark matter (CDM) cosmologies, in which the observed structure builds hierarchically from the bottom–up. In these models, the first object to form are also the smallest; objects with increasingly larger masses assemble subsequently by accretion and mergers.

Although the first object are beyond the redshift limit ($z \sim 5$) of direct detection by present–day instruments, CDM models predict that they formed not exceedingly far from this limit, during the interval ($10 \lesssim z \lesssim 20$). Observations of the intergalactic medium (IGM) reveal that it was both highly ionized, and enriched with heavy elements, before redshift $z \sim 5$, supporting the notion that astrophysical objects exist beyond the current $z \sim 5$ horizon. In this thesis, we were motivated by these observations and theoretical predictions, to study various aspects of non–gravitational physics to estimate when and how the first generation of astrophysical objects formed, and to examine some of their properties. We then addressed the question of whether we could observe these objects in the foreseeable future, or at least infer their existence indirectly from their effects on their surroundings.

The first three chapters were devoted to theoretical considerations. In Chapter 2, we simulated the collapse of spherically symmetric clouds, and derived the minimum mass of clouds that could cool efficiently via H_2. We found that H_2 cooling is important only in the central, shock–heated regions, but showed that primordial clouds as small as $10^3 M_\odot$ are able to collapse at high redshifts, due to the gravitational potential of the cold dark matter component. In Chapters 3 and 4, we examined the effect of ultraviolet (UV) photons on the H_2 abundance in homogeneous clouds. We found that the effect depends primarily on the gas density. At densities expected in the $z \lesssim 30$ universe, we found the H_2 molecules to be fragile, and easily photodissociated by the small UV flux of the first few stars. On the other hand, we showed that at very high densities, UV photons can have the opposite effect, and trigger an increase in the H_2 abundance.

The remaining four chapters were devoted to connections with observations. In Chapters 5 and 6, we examined several signatures of an early generation of stars, using the IGM metallicity to calibrate the high–redshift star formation rates. We concluded that two of these signatures, the suppression of the CMB anisotropies due to reionization, and the spectral distortion of the CMB due to processing of starlight by stardust, are likely to be observed with the next generation of instruments. Most importantly, we showed that the Next Generation Space Telescope could directly image numerous $z > 10$ star–clusters. In Chapter 7, we address the formation and signatures of high redshift quasars, by extrapolating the observed quasar luminosity function to high redshifts and low luminosities. Calculations similar to those for stars in Chapters 5 and 6 showed that the quasar population can have effects comparable to those stars. Finally, in Chapter 8, we confronted quasar formation models with the deepest observations available, the Hubble Deep Field, and showed that we use this data to place useful constraints on our models.

We currently have data on the state of the universe when its age was either less than million years or more than a billion years. The work in this thesis focused on the important transition epoch during which the first astrophysical objects, either small clusters of stars, or low–luminosity quasars, appeared. By quantifying some of the properties of these objects, and their signatures expected to be observed by forthcoming satellite missions, we hope that this thesis has contributed to our picture of structure formation in the universe.

APPENDICES

APPENDIX A: HEATING AND COOLING TERMS

The heating $\Lambda_{\rm heat}$ is given (in units of ergs s^{-1} cm^{-3}) by a sum of seven terms, corresponding to reactions (22)-(28) from Appendix B. All of these have the identical form

$$\Lambda_{\rm heat,i} = 4\pi n_{\rm i} \int d\nu \sigma_{\nu,\rm i} J_\nu \frac{(\nu - \nu_{\rm th,i})}{\nu} f_{\rm heat,\nu}.$$

Here $n_{\rm i}$ is the number density of the dissociated species in cm^{-3}, $\sigma_{\nu,\rm i}$ is the cross-section in cm^2, J_ν is the radiation intensity in erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}, $\nu_{\rm th,i}$ is the threshold frequency for the photon absorption in Hz, and $f_{\rm heat,\nu}$ is a factor between 0.2 and 1, taking into account of the fact that only fraction of the energy of a photo-electron is thermalized between all the species (Voit 1991), the rest of the energy being exhausted on excitations and ionizations/dissociations.

The cooling $\Lambda_{\rm cool}$ is given by the sum of the 12 terms below. In Chapter 2, we used the thermal, and in Chapter 3, the non-thermal Compton scattering formula. References are Black (1981) and Cen (1992) for the first 10 terms, Spitzer & Hart (1971) for the Gaunt factor $g_{\rm ff}$ appearing in $\Lambda_{\rm Brem}(T)$, and Lepp & Shull (1983) for $\Lambda_{\rm H_2}(T)$. We used equation 24.42 in Peebles (1993) to derive $\Lambda_{\rm Comp}(T)$. In these formulas, T denotes the gas temperature in K, and n_i the number densities in cm^{-3}.

Collisional ionization of H, He, *and* He$^+$*:*

$$\Lambda_{\rm H}(T) = 1.27 \times 10^{-21} T^{\frac{1}{2}} (1 + T_5^{\frac{1}{2}})^{-1} \exp(-157809.1/T) n_{\rm e} n_{\rm H}$$
$$\Lambda_{\rm He}(T) = 9.38 \times 10^{-22} T^{\frac{1}{2}} (1 + T_5^{\frac{1}{2}})^{-1} \exp(-285335.4/T) n_{\rm e} n_{\rm He}$$
$$\Lambda_{\rm He^+}(T) = 4.95 \times 10^{-22} T^{\frac{1}{2}} (1 + T_5^{\frac{1}{2}})^{-1} \exp(-631515/T) n_{\rm e} n_{\rm He^+}$$

Recombination to H, He, *and* He$^+$*:*

$$\Lambda_{\rm H^+}(T) = 8.70 \times 10^{-27} T^{\frac{1}{2}} T_3^{-0.2} (1 + T_6^{0.7})^{-1} n_{\rm e} n_{\rm H^+}$$
$$\Lambda_{\rm He^+}(T) = 1.55 \times 10^{-26} T^{0.3647} n_{\rm e} n_{\rm He^+}$$
$$\Lambda_{\rm He^{++}}(T) = 3.48 \times 10^{-26} T^{\frac{1}{2}} T_3^{-0.2} (1 + T_6^{0.7})^{-1} n_{\rm e} n_{\rm He^{++}}$$

Dielectric Recombination to He*:*

$$\Lambda_{\mathrm{He^+}}(T) = 1.24 \times 10^{-13} T^{-1.5} \exp(-470000/T)(1 + 0.3\exp(-94000/T)) n_{\mathrm{e}} n_{\mathrm{He^+}}$$

Collisional excitation of H, *and* $\mathrm{He^+}$*:*

$$\Lambda_{\mathrm{H}}(T) = 7.50 \times 10^{-19} (1 + T_5^{\frac{1}{2}})^{-1} \exp(-118348/T) n_{\mathrm{e}} n_{\mathrm{H}}$$

$$\Lambda_{\mathrm{He^+}}(T) = 5.54 \times 10^{-17} T^{-0.397} (1 + T_5^{\frac{1}{2}})^{-1} \exp(-473638/T) n_{\mathrm{e}} n_{\mathrm{He^+}}$$

Bremsstrahlung:

$$\Lambda_{\mathrm{Brem}}(T) = 1.42 \times 10^{-27} g_{\mathrm{ff}} T^{\frac{1}{2}} n_{\mathrm{e}} (n_{\mathrm{H^+}} + n_{\mathrm{He^+}} + 4 n_{\mathrm{He^{++}}})$$

$$g_{\mathrm{ff}} = 1.10 + 0.34 \exp(-(5.50 - \log T)^2/3.0)$$

Compton cooling:

$$\Lambda_{\mathrm{Comp}}(T) = 1.017 \times 10^{-37} T_\gamma^4 (T - T_\gamma) n_{\mathrm{e}} \text{ (thermal)}$$

$$\Lambda_{\mathrm{Comp}}(T) = 5.64 \times 10^{-54} (T \int J_\nu d\nu - \frac{h}{4k_{\mathrm{B}}} \int J_\nu \nu d\nu) n_{\mathrm{e}} \text{ (non–thermal)}$$

Cooling due to $\mathrm{H_2}$ *molecules:*
A lengthy expression for the term $\Lambda_{\mathrm{H_2}}(T)$ was adopted from Lepp & Shull (1983). The expression in this reference is an analytical fit to the authors' numerical results. It includes cooling from the following mechanisms involving $\mathrm{H_2}$:

- dissociation of $\mathrm{H_2}$ molecules from $\mathrm{H_2} - \mathrm{H_2}$ collisions,
- dissociation of $\mathrm{H_2}$ molecules from $\mathrm{H_2} - \mathrm{H}$ collisions,
- de-excitation from excited rotational and vibrational states, due to $\mathrm{H_2} - \mathrm{H_2}$ collisions,
- de-excitation from excited rotational and vibrational states, due to $\mathrm{H_2} - \mathrm{H}$ collisions.

However, radiative cooling by $\mathrm{H_2}$ is not possible thermodynamically when $T < T_\gamma$; instead $\mathrm{H_2}$ is a source of heating under this condition. In order to correct for the presence of radiative heating from the inverse of the above processes, we multiply the quoted value $\Lambda_{\mathrm{H_2}}(T)$ by the factor $(T - T_\gamma)/(T + T_\gamma)$. Note that there is a typographical error in Lepp & Shull 1983 (confirmed by S. Lepp, private communication), which changes the value of $\Lambda_{\mathrm{H_2}}(T)$ considerably. In their equation (8) on p. 581, 1.10×10^{-13} should be replaced by 1.10×10^{-18}.

APPENDIX B: REACTION RATES AND CROSS SECTIONS

	Reaction	Rate Coefficient (cm^3 sec^{-1})	Reference
(1)	$H + e^- \rightarrow H^+ + 2e^-$	$5.85 \times 10^{-11} T^{\frac{1}{2}} \exp(-157809.1/T)(1+T_5^{\frac{1}{2}})^{-1}$	[2]
(2)	$He + e^- \rightarrow He^+ + 2e^-$	$2.38 \times 10^{-11} T^{\frac{1}{2}} \exp(-285335.4/T)(1+T_5^{\frac{1}{2}})^{-1}$	[1]
(3)	$He^+ + e^- \rightarrow He^{++} + 2e^-$	$5.68 \times 10^{-12} T^{\frac{1}{2}} \exp(-631515.0/T)(1+T_5^{\frac{1}{2}})^{-1}$	[2]
(4)	$H^+ + e^- \rightarrow H + h\nu$	$8.40 \times 10^{-11} T^{-\frac{1}{2}} T_3^{-0.2}(1+T_6^{0.7})^{-1}$	[2]
(5)	$He^+ + e^- \rightarrow He + h\nu$	see expression in reference	[2]
(6)	$He^{++} + e^- \rightarrow He^+ + h\nu$	$3.36 \times 10^{-10} T^{-\frac{1}{2}} T_3^{-0.2}(1+T_6^{0.7})^{-1}$	[2]
(7)	$H + H^+ \rightarrow H_2^+ + h\nu$	see expression in reference	[10]
(8)	$H_2^+ + H \rightarrow H_2 + H^+$	6.40×10^{-10}	[7]
(9)	$H + e^- \rightarrow H^- + h\nu$	$5.57 \times 10^{-17} T^{\frac{1}{2}}$	[9]
(10)	$H + H^- \rightarrow H_2 + e^-$	1.30×10^{-9}	[4]
(11)	$H_2^+ + e^- \rightarrow 2H$	$1.68 \times 10^{-8}(T/300)^{-0.29}$	[8]
(12)	$H_2^+ + H^- \rightarrow H_2 + H$	$5.00 \times 10^{-6} T^{-\frac{1}{2}}$	[3]
(13)	$H^- + H^+ \rightarrow 2H$	$7.00 \times 10^{-7} T^{-\frac{1}{2}}$	[3]
(14)	$H_2 + e^- \rightarrow H + H^-$	$2.70 \times 10^{-8} T^{-\frac{3}{2}} \exp(-43000/T)$	[5]
(15)	$H_2 + H \rightarrow 3H$	see expression in reference	[11]
(16)	$H_2 + H_2 \rightarrow H_2 + 2H$	see expression in reference	[11]
(17)	$H_2 + H^+ \rightarrow H_2^+ + H$	$2.4 \times 10^{-9} \exp(-21200/T)$	[13]
(18)	$H_2 + e^- \rightarrow 2H + e^-$	$4.38 \times 10^{-10} \exp(-102000/T) T^{0.35}$	[13]
(19)	$H^- + e^- \rightarrow H + 2e^-$	$4.00 \times 10^{-12} T \exp(-8750/T)$	[13]
(20)	$H^- + H \rightarrow 2H + e^-$	$5.30 \times 10^{-20} T^{2.17} \exp(-8750/T)$	[13]
(21)	$H^- + H^+ \rightarrow H_2^+ + e^-$	see expression in reference	[13]
		Cross-section (cm^2)	
(22)	$H + h\nu \rightarrow H^+ + e^-$	$\sigma(\nu) = a[\beta(\frac{\nu}{\nu_T})^{-s} + (1-\beta)(\frac{\nu}{\nu_T})^{-s-1}]$ $\nu \geq \nu_T$	[12]
(23)	$He + h\nu \rightarrow He^+ + e^-$	"	[12]
(24)	$He^+ + h\nu \rightarrow He^{++} + e^-$	"	[12]
(25)	$H^- + h\nu \rightarrow H + e^-$	"	[12]
(26)	$H_2^+ + h\nu \rightarrow H + H^+$	$\sigma(\nu, T) = (p_1 + p_2\nu)\exp((\nu - p_3)^2/2p_4^2)$	[14]
(27)	$H_2 + h\nu \rightarrow H_2^+ + e^-$	see expression in reference	[13]
(28)	$H_2 + h\nu \rightarrow 2H$	$\sigma(\nu) = 3 \times 10^{-22}$ for 11.26 eV $< h\nu <$ 13.6 eV	[6]

1. Black (1978); 2. Cen (1992); 3. Dalgarno & Lepp (1987); 4. de Jong (1972); 5. Hirasawa (1969); 6. Allison & Dalgarno (1970), and Dalgarno & Stephens (1970); 7. Karpas et al. (1979); 8. Nakashima et al. (1987); 9. Rawlings (1988); 10. Rawlings et al. (1993); 11. Lepp & Shull (1983); 12. Osterbrock (1974); 13. Shapiro & Kang (1987); 14. Stancil (1994).

The table above lists the chemical reactions included in equations (3.1) in § 2. We attempted to include all possibly important reactions. The rates for reactions (5), (7), (15), (16) and (21) and the cross-section for reaction (27) are given by more lengthy expressions, which can be found in the references. The expression for the cross-section in reaction (26) is a fit to the data from Stancil (1994), with temperature-dependent coefficients p_i. The coefficients appearing in the cross-sections for reactions (22)-(25) can be found in Osterbrock (1974).

H_2 *destruction by the Solomon process*

H_2-molecules are dissociated by photons with energies below 13.6eV by a two-step process, as first suggested by Solomon in 1965 (cf. Field et al. 1966) and later shown quantitatively by Stecher & Williams (1967). We calculated the average cross-section for this two-step photo-dissociation between 11.26eV and 13.6eV (reaction 28), by summing the oscillator strengths for the Lyman and Werner bands of H_2 from Allison and Dalgarno (1970), multiplied by the dissociation fractions (for the second step in the two-step process) from Dalgarno and Stephens (1970). We included the self-shielding of the Lyman and Werner lines using the self-shielding function from de Jong, Dalgarno & Boland (1980), assuming thermal broadening of the lines. Note that the optical depth is proportional to $N_{\rm H_2}/\delta v$, therefore increasing δv is equivalent to decreasing $N_{\rm H_2}$.

Fast photo-electrons

In addition to the 28 reactions in the table above, we included the dissociation of H_2, H, and He by collisions with energetic photo-electrons. The rate we used for this dissociation is given in each case by

$$k_{\rm i} = 4\pi n_{\rm He} \int d\nu \left(\frac{\sigma_{\nu,\rm He} J_\nu}{h\nu}\right) \times f_{{\rm dis},\nu} \times \left(\frac{n_{\rm i}\sigma_{\nu,\rm i}}{\sum_{\rm j} n_{\rm j}\sigma_{\nu,\rm j}}\right) \times \left(\frac{h\nu_{\nu,\rm i} - h\nu_{\rm th,i}}{h\nu_{\rm th,i}}\right).$$

The first term under the integral is the photoionization rate of Helium - this is the process that dominates the supply of energetic photo-electrons. The second term ($f_{{\rm dis},\nu}$) gives the fraction of the electron energy spent on dissociations (from Voit 1991), and the third term gives the distribution of these dissociations between the three species H_2, H, and He. The last term converts the energy available for dissociations to the number of dissociations of a given species.

Self-absorption of infrared photons

The H_2-cooling could, in principle, be suppressed if the gas cloud were optically-thick to its emitted infrared-line photons. H_2-cooling occurs mainly due to radiative de-excitation from excited rotational and vibrational levels of the electronic ground state. We obtained cross-sections for these transitions from the transition probabilities given in Turner et al. 1977, and found them to be smaller than $\sim 10^{-26}$ cm^2. Given these small cross-sections and the typically small H_2-fraction ($\lesssim 0.01$), we conclude that an exceedingly high H-column density, $\sim 10^{28}$ cm^{-2}, is necessary to absorb the IR cooling photons.

REFERENCES

Adams, F. C., Freese, K., Levin, J., McDowell, J. 1989, ApJ, 344, 24
Aghanim, N., Désert, F. X., Puget, J. L., & Gispert, R. 1996, A&A, 311, 1
Alcock, C., et al. 1997a, ApJ, 479, 119
Alcock, C., et al. 1997b, ApJ, 486, 697
Allison, A. C. & Dalgarno, A. 1970, Atomic Data, 1, 289
Anninos, P., Zhang, Y., Abel, T., Norman, M. L. 1997, New Ast., 2, 209
Arons, J., & Wingert, D. W. 1972, ApJ, 177, 1
Babul, A., & Ferguson, H. C. 1996, ApJ, 458, 100
Babul, A., & Rees, M. J. 1992, MNRAS, 255, 346
Bajtlik, S., Duncan, R. C., & Ostriker, J. P. 1988, ApJ, 327, 570
Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A.S. 1986, ApJ, 304, 15 (BBKS)
Barnes, J, & Efstathiou, G. 1987, ApJ, 319, 575
Barnes, J., Hut, P., & Goodman, J. 1986, ApJ, 300, 112
Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G. 1997, preprint astro-ph/9703111
Bechtold, J. 1994, ApJS, 91, 1
Bechtold, J., Crotts, A. P. S., Duncan, R. C., & Fang, Y. 1994, ApJ, 437, L83
Bennett, C. L., Banday, A. J., Gorski, K. M., Hinshaw, G., Jackson, P., Keegstra, P., Kogut, A., Smoot, G. F., Wilkinson, D. T., & Wright, E. L. 1996, ApJ, 464, L1
Bennett, C. L., et al. 1996, ApJ, 464, L1
Bertschinger, E. 1985, ApJS, 58, 39
Binney, J., & Tremaine, S. 1987, Galactic Dynamics, Princeton Univ. Press, Princeton, pp. 287-291
Birkhoff, G. D. 1923, Relativity and Modern Physics, Harvard Univ. Pres, Cambridge, Mass
Black, J.H. 1978, ApJ, 222, 125
Black, J.H. 1981, MNRAS, 197, 553
Bond, J. R. 1995a, in Cosmology and Large Scale Structure, ed. R Schaeffer, Elsevier Science Publishers, The Netherlands
Bond, J. R. 1995b, in Theory and Observations of the Cosmic Microwave Background Radiation, ed. Schaeffer, R., Elsevier Science Publishers, The Netherlands
Bond, J. R., Carr, B. J., & Hogan, C. J. 1991, ApJ, 367, 420
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Bond, J. R., & Efstathiou, G. 1987, MNRAS, 226, 665
Bond, J. R., & Myers, S. 1996, ApJS, 103, 63
Bonnor, W. B. 1957, MNRAS, 117, 111
Boyle, B. J., Jones, L. R., Shanks, T., Marano, B., Zitelli, V., & Zamorani, G. 1991, in Crampton, D., ed., ASP Conf. Series No. 21, The Space Distribution of Quasars, Astron. Soc. Pacif., San Francisco, p. 191
Boyle, B. J., Shanks, T., & Peterson, B. A., 1988, MNRAS, 235, 935
Bruzual, G., & Charlot, S. 1996, in preparation. The models are available from the anonymous ftp site gemini.tuc.noao.edu.
Bunn, E. F., Scott, D., & White, M. 1995, ApJ, 441, L9
Bunn, E. F., & White, M. 1997, ApJ, 480, 6
Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582

Carlberg, R. G. 1981, MNRAS, 197, 1021
Carlberg, R. G. 1990, ApJ, 350, 505
Carr, B. J., Bond, J. R., & Arnett, W. D. 1984, ApJ, 277, 445
Carroll, S. M., Press, W. H., Turner, E. L. 1992, ARA&A, 30, 499
Cen, R. 1992, ApJ, 78, 341
Cen, R., Miralda-Escude, J., Ostriker, J. P., & Rauch, M. 1994, ApJ, 437, L9
Cen, R., & Ostriker, J. P. 1993, ApJ, 414, 507
Cen, R., & Ostriker, J. P. 1994, ApJ, 431, 451
Charlot, S., & Fall, M. S. 1991, ApJ, 378, 471
Cheney, J. E., & Rowan–Robinson, M. 1981, MNRAS, 195, 831
Cole, S., Aragón-Salamanca, A, Frenk, C., Navarro, J., Zepf, S. 1994, MNRAS 271, 781
Conti, A., Kennefick, J. D., Martini, P., & Osmer, P. S. 1998, in preparation
Copi, C. J., Schramm, D. N., & Turner, M. S. 1995, Science, 267, 192
Couchman, H. M. P. & Rees, M. J. 1986, MNRAS, 221, 53
Cowie, L. 1996, in HST and the High Redshift Universe, Proc. 37th Herstmonceux conference, eds. Tanvir, N. R., Aragón-Salamanca, A., Wall, J.V., World Scientific, Singapore, preprint astro-ph/9609158
Cowie, L. L., Songaila, A., Kim, T.-S., & Hu, E. M. 1995 AJ, 109, 1522
Dalcanton, J. J., Canizares, C. R., Granados, A., Steidel, C. C., & Stocke, J. T. 1994, ApJ, 424, 550
Dalgarno, A. & Stephens, T. L. 1970, ApJ, 160, L107
Dalgarno, A., & Lepp, S. 1987, in Astrochemistry: Proceedings of IAU Symposium No. 120, ed. M. S. Varya & S. P. Tarafdar, Reidel Publishing Co., Dordrecht, p. 109
Davidsen, A. F., Kriss, G. A., & Zheng, W. 1996, Nature, 380, 47
de Jong, T. 1972, A&A, 20, 263
de Jong, T., Dalgarno, A., & Boland, W. 1980, A&A, 91, 68
Dekel, A., & Rees, M. J. 1987, Nature, 326, 455
Dekel, A., & Silk, J. 1986, ApJ, 303, 39
Dey, A., Spinrad, H., Stern, D., Graham, J. R., Chaffee, F. H. 1998, ApJL, in press, preprint astro-ph/9803137
Dinshaw, N., Foltz, C. B., Impey, C. D., Weymann, R. J., & Morris, S. L. 1995, Nature, 373, 223
Dinshaw, N., Impey, C. D., Foltz, C. B., Weymann, R. J.,
Doroshkevich, A.G., Zel'dovich, Ya. B., & Novikov, I. D. 1967, Sov. Astr. A. J., 11, 233.
Dove, J. B., & Shull, J. M. 1994, ApJ, 430, 222
Draine, B. T., & Lee, H. M. 1984, ApJ, 285, 89
Dubinski, J., Mihos, J. C., Hernquist, L. 1996, ApJ, 462, 576
Dwek, E. 1988, ApJ, 329, 814
Dwek, E. 1997, in Science with the Next Generation Space Telescope, Astr. Soc. of the Pacific Conference Series, ed. E. P. Smith & A. Koratkar
Efstathiou, G. 1992, MNRAS, 256, 43p
Efstathiou, G., Bond, J.R., & White, S. D. M. 1992, MNRAS, 258, 1p
Efstathiou, G., & Bond, J. R. 1987, MNRAS, 227, 33p
Efstathiou, G., & Rees, M. J. 1988, MNRAS, 230, 5p
Eisenstein, D. J. 1997, ApJ, submitted, preprint astro-ph/9709054
Eisenstein, D.J., & Loeb, A. 1995a, ApJ, 439, 520
Eisenstein, D.J., & Loeb, A. 1995b, ApJ, 443, 11
Elston, R. A. W., Santiago, B. X., & Gilmore, G. F. 1996, New Astronomy, 1, 1
Elvis, M., Wilkes, B. J., McDowell, J. C., Green, R. F., Bechtold, J., Willner, S. P., Oey, M. S., Polomski, E., & Cutri, R. 1994, ApJS, 95, 1
Fabian, A. C. & Barcons, X. 1992, ARA&A, 30, 429
Fan, X.M. 1994, PhD Thesis, Columbia Univ., New York
Fang, Y., Duncan, R. C, Crotts, A. P. S., & Bechtold, J. 1996, 462, 77
Field, G. B., Somerville, W. B., and Dressler, K. 1966, ARA&A 4, 207
Fillmore, J. A., & Goldreich, P. 1984, ApJ, 281, 1
Fixsen, D. J., Cheng, E. S., Gales, J. M., Mather, J. C., Shafer, R. A., Wright, E. L. 1996, ApJ,

473, 576
Fixsen, D. J., Dwek, E., Mather, J. C., Bennett, C. L., & Shafer, R. A. 1998, ApJ, in press, preprint astro-ph/9803021
Fixsen, D. J., et al. 1996, ApJ, 473, 576
Flynn, C., Gould, A., & Bahcall, J. N. 1996, ApJL, 466, 55
Frank, J., King, A. R., & Raine, D. J. 1992, Accretion Power in Astrophysics, 2nd ed., Cambridge Univ. Press, Cambridge
Franx, M, Illingworth, G. D., Kelson, D. D., van Dokkum, P. G., & Tran, K.-V. 1997, ApJ, 486, L75
Fukugita, M., & Kawasaki, M. 1994, MNRAS, 269, 563
Gamow, G. 1948, Nature, 162, 680
Gelb, J. M., & Bertschinger, E. 1994, ApJ, 436, 467
Geller, M. J., & Huchra, J. P. 1989, Science, 246, 897
Giroux, M. L., Fardal, M. A., & Shull, J. M. 1995, ApJ, 451, 477
Giroux, M. L., & Shapiro, P. R. 1996, ApJS, 102, 191
Giroux, M. L., & Shull, J. M. 1997, AJ, 113, 1505
Gnedin, N. Y., 1998, MNRAS, 294, 407
Gnedin, N. Y., & Ostriker, J. P. 1997, ApJ, 486, 581
Gott, J.R. 1975, ApJ, 201, 296
Gould, A. 1995, ApJ, 455, 37
Governato, F., Baugh, C. M., Frenk, C. S., Cole, S., Lacey, C. G., Quinn, T., & Stadel, J. 1998, Nature, 392, 359
Gunn, J. E., & Peterson, B. A., 1965, ApJ, 142, 1633
Guth, A. H. 1981, Phys. Rev. D, 23, 347
Haardt, F., Madau, P. 1996, ApJ, 461, 20
Haehnelt, M. G., Natarajan, P., & Rees, M. J. 1998, MNRAS, submitted, preprint astro-ph/9712259
Haehnelt, M. G., & Rees, M. J. 1993, MNRAS, 263, 168
Haiman, Z., Madau, P., & Loeb, A. 1998, in preparation
Haiman, Z., Rees, M. J., & Loeb, A. 1996, ApJ, 467, 522 (HRL96)
Haiman, Z., Rees, M. J., & Loeb, A. 1997, ApJ, 476, 458 (HRL97)
Haiman, Z., Thoul, A., & Loeb, A. 1996, ApJ, 464, 523 (HTL96)
Haiman, Z., & Loeb, A. 1997, ApJ, 483, 21 (HL97)
Haiman, Z., & Loeb, A. 1998a, in Science with the Next Generation Space Telescope, eds. E. Smith & A. Koratkar, preprint astro-ph/9705144
Haiman, Z., & Loeb, A. 1998b, ApJ, in press, astro-ph/9710208
Harrison, E. R. 1970, Phys. Rev. D, 2726, 1
Hartwick, F. D. A., & Schade, D. 1990, ARA&A, 28, 437
Hellsten, U., Davé, R., Hernquist, L., Weinberg, D. H., & Katz, N. 1997, ApJ, 487, 482
Hernquist, L., Katz, N., Weinberg, D. H., & Miralda-Escudé, J. 1996, ApJ, 457, L51
Hernquist, L., Katz, N., & Weinberg, D. 1995, ApJ, 442, 57
Hindmarsh, A.C. 1983, in Scientific Computing, ed. R.S. Stepleman et al. , North-Holland, Amsterdam, pp. 55-64
Hirasawa, T. 1969, Prog. Theor. Phys., v. 42, no. 3, p. 523
Hogan, C. J., Anderson, S. F., Rugers, M. H. 1997, AJ, 113, 1495
Hollenbach, D., & McKee, C.F. 1979, ApJS, 41, 555
Hu, W., & White, M. 1997, ApJ, 479, 568
Hutchins, J. B. 1976, ApJ 205, 103
Ikeuchi, S. 1986, Ap&SS, 118, 509
Jacobsen, P., Boksenberg, A., Deharveng, J. M., Greenfield, P., Jedrzejewski, R., & Paresce, F. 1994, Nature, 370, 35
Jeans, J. H. 1928, Astronomy and Cosmogony, Cambridge Univ. Press, Cambride, UK
Kamionkowski, M., Spergel, D. N., & Sugiyama, N. 1994, ApJL, 426, L57
Karpas, Z., Anicich, V., and Huntress, W.T. 1979, J. Chem. Phys., 70, 2877
Kashlinsky, A., & Rees, M.J. 1983, MNRAS, 205, 955

Katz, N., Quinn, T., & Gelb, J. M. 1993, MNRAS, 265, 689
Kauffmann, G., Guiderdoni, B., & White, S. D. M. 1994, MNRAS, 267, 981
Kauffmann, G., & White, S. D. M. 1993, MNRAS, 261, 921
Klypin, A., Holtzman, J., Primack, J., & Regős, E. 1993, ApJ, 416, 1
Kogut, A. 1996, in Proceedings from XVI Moriond Astrophysics meeting, Les Arcs, France, preprint astro-ph/9607100
Kolb, E.W., & Turner, M.S. 1990, The Early Universe, Addison-Wesley, Redwood City, CA
Kormendy, J., Bender, R., Magorrian, J., Tremaine, S., Gebhardt, K., Richstone, D., Dressler, A., Faber, S. M., Grillmair, C., & Lauer, T. R. 1997, ApJ, 482, L139
Kulkarni, V. P., & Fall, S. M. 1993, 413, L63
Kurucz, R. 1993, CD-ROM No. 13, ATLAS9 Stellar Atmosphere Programs
Lacey, C. G., & Cole, S. 1993, MNRAS, 262, 627
Lacey, C. G., & Cole, S. 1994, MNRAS, 271, 676
Lahav, O. 1986, MNRAS, 220, 259
Larson, R. B. 1974, MNRAS, 166, 585
Larson, R. B. 1975, MNRAS, 173, 671
Larson, R. B. 1986a, MNRAS, 218, 419
Larson, R. B. 1986b, in Norman, C. A., Renzini, A., Tosi, M., eds., Stellar Populations, Cambridge Univ. Press, Cambridge, p. 125
Larson, R.B. 1969, MNRAS, 145, 271
Leitherer, C., Ferguson, H. C., Heckman, T. M., & Lowenthal, J. D. 1995, ApJ, 454, L19
Lemaitre, G. 1933, Ann. Soc. Sci. Bruxelles, A53, 51
Lepp, S., & Shull, J. M. 1983, ApJ, 270, 578
Lepp, S., & Shull, J. M. 1984, ApJ, 280, 465
Liddle, A. R., & Lyth, D. 1992, Phys. Let., B291, 391
Lifshiftz, E. M. 1946, J. Phys., 10, 116
Lin, D. N. C., & Murray, S. D. 1992, ApJ, 394, 523
Linde, A. D. 1982, Phys. Lett., 108B, 389
Loeb, A. 1993, ApJ, 403, 542
Loeb, A. 1996, ApJ, 459, L5
Loeb, A. 1998, in Science with the Next Generation Space Telescope, eds. E. Smith & A. Koratkar, preprint astro-ph/9704290
Loeb, A., & Rasio, F. A. 1994, ApJ, 4322, 52
Loeb, A., & Haiman, Z. 1997, ApJ, 490, 571 (LH97)
Lowenthal, J. D., Koo, D. C., Guzman, R., Gallego, J., Phillips, A. C., Faber, S. M., Vogt, N. P., Illingworth, G. D., & Gronwall, C. 1997, ApJ, 481, 673
Lu, L., Sargent, W. L. W., Barlow, T. A., Churchill, C. W., Vogt, S. S. 1996, ApJS, 107, 475
Lu, L., Wolfe, A. M., & Turnshek, D. A. 1991, ApJ, 367, 19
Lucy, L. B., Danziger, I. J., Gouiffes, C., & Bouchet, P. 1991, in Supernovae, ed. S. E. Woosley, Springer, New York, p. 82
Lynden–Bell, D. 1967, MNRAS, 136, 101
Méndez, R. A., Minniti, D., De Marchi, G., Baker, A., & Couch, W. J. 1996, MNRAS, 283, 666
Madau, P. 1995, ApJ, 441, 18
Madau, P., Ferguson, H. C., Dickinson, M. E., Giavalisco, M., Steidel, C. C., & Fruchter, A. 1996, MNRAS, 283, 1388
Madau, P., & Meiksin, A. 1994, ApJ, 433, L53
Madau, P., & Shull, J. M. 1996, ApJ, 457, 551
Maeder, A. 1992, ApJ, 264, 105
Magorrian, J., et al. 1997, submitted to AJ, preprint astro-ph/9708072
Mather, J, & Stockman, P. 1996, STScI Newsletter, v. 13, no. 2, p. 15
Mather, J. C. et al. 1994, ApJ, 420, 439
Mathis, J. S. 1990, ARA&A, 28, 37
Matsuda, T., Sato, H., & Takeda, H. 1969, Prog. Theor. Phys., v. 42, no. 2, p. 219
McLeod, K. 1996, in ESO-IAC conference on Quasar Hosts, Tenerife, in press
Meiksin, A., & Madau, P. 1993, ApJ, 412, 34

Miller, G. E., & Scalo, J. M. 1979, ApJS, 41, 513
Miralda-Escudé, J. 1997, ApJ, in press, preprint astro-ph/9708253
Miralda-Escudé, J., Cen, R., Ostriker, J. P., & Rauch, M. 1996, ApJ, 471, 582
Miralda-Escudé, J., & Ostriker, J. P. 1990, ApJ, 350, 1
Miralda-Escudé, J., & Rees, M. J. 1994, MNRAS, 266, 343
Miralda-Escudé, J., & Rees, M. J. 1997, ApJ, 478, L57
Nakashima, K., Takayi, H., Nakamura, H. 1987, J. Chem. Phys., 86, 726
Narayan, R. 1996, "Advective Disks", in Proc. IAU Colloq. 163 on Accretion Phenomena & Related Outflows, A.S.P. Conf. Series, eds. D. T. Wickramasinghe, L. Ferrario, G. V. Bicknell, in press, preprint astro-ph/9611113
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1995, MNRAS, 275, 56
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Navarro, J. F., & Steinmetz, M. 1997, ApJ, 478, 13
Novikov, I. D. 1964, Sov. Phys. JETP, 19, 467
Osterbrock, D.E. 1974, Astrophysics of Gaseous Nebulae, Freeman & Co, San Francisco, CA
Ostriker, J. P., Vogeley, M. S., & York, D. G. 1990, ApJ, 364, 405
Ostriker, J. P., & Gnedin, N. Y. 1996, ApJ, 472, 603
Ostriker, J. P., & Heisler, J. 1984, ApJ, 278, 1
Ostriker, J. P., & Steinhardt, P. J. 1995, Nature, 377, 600
Padmanabhan, T. 1993, Structure Formation in the Universe, Cambridge University Press, Cambridge
Palla, F., Salpeter, E. E., & Stahler, S.W. 1983, ApJ, 271, 632
Park, M.-G., Ostriker, J. P. 1989, ApJ, 347, 679
Peacock, J. A., & Heavens, J. A. 1990, MNRAS, 243, 133
Peebles, P. J. E. 1965, ApJ, 142, 1317
Peebles, P. J. E. 1967, ApJ, 147, 859
Peebles, P. J. E. 1968, Nature, 220, 237
Peebles, P. J. E. 1980, Large-Scale Structure of the Universe, Princeton Univ. Press
Peebles, P. J. E. 1982, ApJ, 263, L1
Peebles, P. J. E. 1993, Principles of Physical Cosmology, Princeton University Press, Princeton
Peebles, P.J.E., & Dicke, R.H. 1968, ApJ, 154, 891
Pei, Y. C. 1995, ApJ, 438, 623
Pettini, M., Steidel, C. C., Adelberger, K. L., Kellogg, M., Dickinson, M. E., Giavalisco, M. 1997, to appear in 'ORIGINS', ed. J.M. Shull, C.E. Woodward, and H. Thronson, ASP Conference Series, preprint astro-ph/9708117
Press, W. H., & Schechter, P. L. 1974, ApJ, 181, 425
Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. 1992, Numerical Recipes in Fortran, 2^{nd} ed., Cambridge University Press, Cambridge, UK
Press, W.H., & Gunn, J.E. 1973, ApJ, 185, 397
Quinn, T., Katz, N., & Efstathiou, G. 1996, MNRAS, 278, 49
Rauch, M., Haehnelt, M. G., & Steinmetz, M. 1997, ApJ, 481, 601
Rawlings, J. M. C. 1988, MNRAS, 232, 507
Rawlings, J. M. C., Drew, J.E., Barlow, M.J. 1993, MNRAS, 265, 968
Rees, M. J. 1996, preprint astro-ph/9608196
Rees, M. J., & Ostriker, J. P. 1977, MNRAS, 179, 541
Reimers, D., Köhler, S., Wisotzki, L., Groote, D., Rodriguez-Pascual, P., & Wamsteker, W. 1997, A&A, 326, 489
Renzini, A., & Voli, M. 1981, A&A, 94, 175
Richstone, D., Loeb, A., & Turner, E.L. 1992, ApJ, 393, 477
Rowan–Robinson, M. 1995, MNRAS, 272, 737
Ruzmaikina, T.V. 1973, Sov. Astron., v. 16, no. 6, p. 991
Rybicki, G. B., & Lightman, A. P. 1979, Radiative Processes in Astrophysics, Wiley, New York
Salpeter, E. E. 1964, ApJ, 140, 496
Sasaki, S. 1994, PASJ, 46, 427

Saslaw, W.C., & Zipoy, D. 1967, Nature, 216, 976
Scalo, J. M. 1986, Fundamentals of Cosmic Physics, vol. 11, p. 1-278
Schaller, G., Schaerer, D., Meynet, G., & Maeder, A. 1992, A&ASS, 96, 269
Schneider, D. P., Schmidt, M., & Gunn, J. E. 1991, AJ, 102, 837
Scott, D., Silk, J., & White, M. 1995, Science, 268, 829
Shapiro, P. R., Giroux, M. L., & Babul, A. 1994, ApJ, 427, 25
Shapiro, P. R., & Giroux, M. L. 1987, ApJ, 321, L107
Shapiro, P. R., & Kang, H. 1987, ApJ, 318, 32
Silk, J. 1977, ApJ, 211, 638
Small, T. A., & Blandford, R. D. 1992, MNRAS, 259, 725
Smette, A., Robertson, J. G., Shaver, P. A., Reimers, D., Wisotzki, L., & Köhler, Th. 1995, A&A Suppl., 113, 199
Smith, E. P., & Koratkar, A. eds, 1998, Science with the Next Generation Space Telescope, Ast. Soc. of the Pacific Conf. Series, v. 133
Smoot, G. F. 1992, ApJL, 396, 1
Smoot, G. F., et al. 1992, ApJ, 396, L1
Soltan, A. 1982, MNRAS, 200, 115
Somerville, R. S., Primack, J. R. 1998, MNRAS, submitted
Songaila, A. 1997, ApJ, 490, L1
Songaila, A., & Cowie, L. L. 1996, AJ, 112, 335
Spitzer, L. Jr., & Hart, M. H. 1971, ApJ, 166, 483
Stancil, P.C. 1994, ApJ, 430, 360
Stecher, T. P., & Williams, D. A. 1967, ApJ, 149, L29
Steidel, C. C., Adelberger, K. L., Dickinson, M. E., Giavalisco, M, Pettini, M., Kellogg, M. 1998, ApJ, 492, 428
Steidel, C. C., Giavalisco, M., Dickinson, M., & Adelberger, K. L. 1996, AJ, 112, 352
Steinhardt, P. J., & Turner, M. S. 1984, Phys. Rev. D, 29, 2162
Steinmetz, M. 1995, preprint astro-ph/9502013
Tegmark, M., & Silk, J. 1995, ApJ, 441, 458
Tegmark, M., Silk, J., & Blanchard, A. 1994, ApJ, 420, 484
Tegmark, M., Silk, J., & Blanchard, A. 1995, ApJ, 434, 395
Tegmark, M., Silk, J., Rees, M. J., Abel, T., & Blanchard, A. 1997, ApJ, 474, L1
Thompson, D., Djorgovski, S., & Trauger, J. 1995, AJ, 110, 963
Thoul, A. A., & Weinberg, D. H. 1995, ApJ, 442, 480
Thoul, A. A., & Weinberg, D. H. 1996, ApJ, 465, 608
Truran, J. W., & Cameron, A. G. W. 1971, Ap. Space Sci., 14, 179
Turner, E. L. Ostriker, J. P., & Gott, J., R., III. 1984, ApJ, 284, 1
Turner, J., Kirby-Docken, K., & Dalgarno, A. 1977, ApJS, 35, 281
Tytler, D. & Fan, X.M. 1994, BAAS, 185, 7004
Tytler, D. et al. 1995, in QSO Absorption Lines, ESO Astrophysics Symposia, ed. G. Meylan, Springer, Heidelberg, p.289
van den Hoek, L. B., & Groenwegen, M. A. T. 1997, A&AS, 123, 305
van der Marel, R. de Zeeuw, P. T., Rix, H-W., & Quinlan, G. D. 1997, Nature, 385, 610
Viana, P. T. P., & Liddle, A. R. 1996, MNRAS, 281, 323
Vikhlinin, A. A. 1995, Astron. Lett., vol. 21, no. 3., p. 366
Vogeley, M. S. 1997, ApJ, in press, preprint astro-ph/9711209
Voit, G. M. 1991, ApJ, 158, 377
Warren, S. J., Hewett, P. C., & Osmer, P. S. 1994, ApJ, 421, 41
Weaver, T. A., & Woosley, S. E. 1980, Ann. N.Y. Ac. Sci., 336, 335
Weinberg, D. H., Hernquist, L., & Katz, N. 1997, ApJ, 477, 8
White, M., Scott, D., & Silk, J. 1994, ARA& A, 32, 319
White, S. D. M. 1994, Les Houches Lecture Notes, in press, MPA preprint 831
White, S. D. M., Efstathiou, G., & Frenk, C. S. 1993, MNRAS, 262, 1023
White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341
Williams, R. E., et al. 1996, AJ, 112, 1335

Woltjer, L. 1990, in Active Galactic Nuclei, SAAS-FEE Adv. Course 20, Swiss Society for Astroph. & Astron., eds. T. J. L. Courvoisier & M. Mayor, Springer, Berlin

Woosley, S. E., & Weaver, T. A. 1986, ARA&A, 24, 205

Woosley, S. E., & Weaver, T. A. 1995, A&A Supp., 101, 181

Wright, E. L. 1981, ApJ, 250, 1

Wright, E. L. 1990, ApJ, 353, 411

Wright, E. L., et al. 1994, ApJ, 420, 450

Yi, I. 1996, ApJ, 473, 645

Zaldarriaga, M. 1997, Phys. Rev. D, 55, 1822

Zaldarriaga, M., Spergel, D., & Seljak, U. 1997, ApJ, 488, 1

Zel'dovich, Ya. B. 1965, Adv. Astr. Astroph., 3, 241

Zel'dovich, Ya. B. 1972, MNRAS, 160, 1p

Zel'dovich, Ya. B., & Novikov, I. D. 1964, Sov. Phys. Dokl., 158, 811

Zheng, W., Kriss, G. A., Telfer, R. C., Grimes, J. P., Davidsen, A. F. 1997, ApJ, 475, 469