Fossil H II regions: self-limiting star formation at high redshift

S. Peng Oh1* and Zoltán Haiman2
1Theoretical Astrophysics, Mail Code 130-33, Caltech, Pasadena, CA 91125, USA
2Department of Astronomy, Columbia University, 550 West 120th Street, New York, NY 10027, USA

Accepted 2003 August 8. Received 2003 August 2; in original form 2003 July 8

ABSTRACT
Recent results from the Wilkinson Microwave Anisotropy Probe (WMAP) satellite suggest that the intergalactic medium (IGM) was significantly reionized at redshifts as high as \( z \approx 17 \). At this early epoch, the first ionizing sources probably appeared in the shallow potential wells of mini-haloes with virial temperatures \( T_{\text{vir}} < 10^4 \text{ K} \). Once such an ionizing source turns off, its surrounding H II region Compton cools and recombines. Nonetheless, we show that the ‘fossil’ H II regions left behind remain at high adiabats, prohibiting gas accretion and cooling in subsequent generations of mini-haloes. This greatly amplifies feedback effects explored in previous studies, and early star formation is self-limiting. We quantify this effect to show that star formation in mini-haloes cannot account for the bulk of the electron scattering opacity measured by WMAP, which must be due to more massive objects. We argue that gas entropy, rather than IGM metallicity, regulates the evolution of the global ionizing emissivity and impedes full reionization until lower redshifts. We discuss several important consequences of this early entropy floor for reionization. It reduces gas clumping, curtailing the required photon budget for reionization. An entropy floor also prevents H2 formation and cooling, due to reduced gas densities: it greatly enhances feedback from ultraviolet photodissociation of H2. An early X-ray background would also furnish an entropy floor to the entire IGM; thus, X-rays impede rather than enhance H2 formation. Future 21-cm observations may probe the topology of fossil H II regions.

Key words: galaxies: formation – intergalactic medium – cosmology: theory – large-scale structure of Universe.

1 INTRODUCTION
The high optical depth \( \tau = 0.17 \pm 0.04 \) detected by the Wilkinson Microwave Anisotropy Probe (WMAP) satellite has lent greater credence to the notion of an early period of star formation and reionization, \( z_r = 17 \pm 8 \) (Kogut et al. 2003; Spergel et al. 2003). If indeed the first stars formed at high redshift \( z \sim 20 \), they are expected to form in mini-haloes1 with shallow potential wells, in which H2 cooling is dominant (Abel, Bryan & Norman 2000; Bromm, Coppi & Larson 2002). More massive haloes with \( T_{\text{vir}} > 10^4 \text{ K} \), in which collisional ionization and line cooling can operate, are expected to be very rare at these redshifts. Modulo the effects of ultraviolet (UV) feedback on H2 formation and cooling, stars forming in such haloes could therefore play a dominant role in an early reionization epoch. A great deal of effort has gone into assessing the impact of UV feedback on H2 cooling, and the counter-heating results of positive feedback effects such as an early X-ray background (e.g. Ciardi, Ferrara & Abel 2000; Haiman, Abel & Rees 2000; Machacek, Bryan & Abel 2001; Ricotti, Gnedin & Shull 2001).

The main point of this paper is that the population of mini-haloes is likely to be considerably sparser than previously assumed. This is because mini-halo formation is strongly suppressed even inside the fossil H II regions of dead ionizing sources. Although such H II regions recombine and cool by Compton scattering with cosmic microwave background (CMB) photons, they cannot cool back to the temperature of the undisturbed intergalactic medium (IGM). Strong Jeans mass filtering takes place (Gnedin 2000), and subsequent mini-haloes will no longer be able to accrete gas due to the smoothing effects of finite gas pressure. Thus, once any patch of the Universe is ionized, it can no longer host any more mini-haloes, even if it subsequently cools and recombines. In effect, the birth of the first stars leads to the demise of the mini-halo population: only one generation of stars can form within these shallow potential wells.

While many authors have noted and commented on the Jeans mass filtering in actively ionized regions of the IGM after full reionization (Bullock, Kravtsov & Weinberg 2000; Benson et al. 2002; Somerville 2002), the Jeans filtering in fossil H II regions after ionizing sources have turned off has received little attention. An exception is the recent work by Ricotti, Gnedin & Shull (2001, 2002a,b), who

*E-mail: peng@tapir.caltech.edu
1 For the purposes of this paper, a mini-halo is defined as any halo with \( T_{\text{vir}} < 10^4 \text{ K} \) that cannot cool via atomic line cooling.
consider the effect of a cosmic UV field on the H$_2$ abundance. They find that when an H II front sweeps by a fluid element that has already contracted to a high density, it can result in an overall increase in the H$_2$ abundance after the UV field is turned off. In this paper, we consider the behaviour of the background gas near the mean IGM density inside fossil H II regions around the first ionizing sources. We show that the entropy floor established in the fossil H II regions prevents this gas from contracting and building up dense cores of mini-haloes (on scales much smaller than possible to address in cosmological simulations). We argue that this will strongly suppress the mini-halo population, with the following interesting consequences.

(i) Impact of mini-halo population on reionization. Since the recombination time is shorter than the Hubble time at high redshift, $t_{\text{rec}} \ll t_{\text{H}}$, and the ionizing sources, expected to be massive stars, have short lifetimes, $t_{\text{M}} = 3 \times 10^7$ yr $\ll t_{\text{H}}$, many generations of star formation are required to keep a given patch of the IGM ionized. However, once the first generation of stars born in mini-haloes dies out, subsequent generations will not be able to form in mini-haloes in a previously ionized patch of IGM, even after the patch cools and recombines. Previous semi-analytical work on the reionization history has already included photoionization feedback in actively ionized regions of the IGM (Haiman & Loeb 1997, 1998; Haiman & Holder 2003; Wyithe & Loeb 2003), by excluding new sources in shallow potential wells from forming in these regions. Here we show that the additional exclusion of ionizing sources from the fossil H II regions produces a strong additional modification of the reionization history, both quantitatively and qualitatively (see Fig. 16 below). Specifically, the feedback in fossil H II regions inevitably produces a non-monotonic reionization history with an early peak of partial reionization, followed by recombination and eventual full reionization. This is qualitatively similar to the reionization history derived by Cen (2003) and Wyithe & Loeb (2003), although the physical reason for the non-monotonic evolution in the present work is different (entropy injection, rather than a Population III to II transition caused by a universal metallicity increase). Regardless of the extent of UV feedback effects on H$_2$ production and cooling, the majority of stars which reionized the Universe were hosted by more massive haloes able to survive Jeans mass filtering, $T_{\text{vir}} > 10^4$ K.

(ii) Photon budget for reionization. Dense gas that collects in mini-haloes can form a considerable sink of ionizing photons by boosting the overall effective gas clumping factor $C_{\text{II}} = (n e^+/(n_e e^+))^2$ (Haiman, Abel & Madau 2001; Barkana & Loeb 2002; Shapiro et al. 2003). The clumping factor $C_{\text{II}}$ increases rapidly at low redshift as structure formation progresses; thus, despite the higher mean gas density $n \propto (1+z)^3$ at high redshift, the recombination time $t_{\text{rec}} = 1/(an_{\text{II}} C_{\text{II}})$ does not evolve strongly from $z = 10$ to 20. However, if the Universe is filled up with fossil H II regions at high redshift, the photon budgets required to subsequently ionize it and to keep it ionized are much lower, since gas clumping in mini-haloes is strongly suppressed.

(iii) Source clustering. Since pre-heating boosts the threshold mass required for efficient gas cooling and star formation, it increases the mean bias of the early protogalaxy population by a factor of a few (e.g. see Fig. 2 of Oh, Cooray & Kamionkowski 2003). This is likely to increase the clustering amplitude of background fluctuations due to these faint unresolved early protogalaxies, such as the free–free background due to ionized haloes (Oh 1999; Oh & Mack 2003), Sunyaev–Zel’dovich fluctuations due to high-redshift H II regions and supernovae (SNe) winds (Knox et al. 1998; Oh et al. 2003; Santos et al. 2003), and the infrared background due to stellar emission (Santos, Bromm & Kamionkowski 2002; Magliocchetti, Salvaterra & Ferrara 2003; Haiman, Spergel & Turner 2003). The increase in clustering bias boosts the amplitude of rms fluctuations by a factor of several. This may be necessary, for instance, if thermal SZ fluctuations from high-redshift supernovae are to account for the small-scale CMB anisotropies observed by BIMA and CBI (Oh et al. 2003).

(iv) Star formation history/metallicity/initial mass function. Initially, each mini-halo is expected to harbour a single massive, metal-free star (no fragmentation is seen in numerical simulations by Abel et al. (2000) and Bromm et al. (2002). Because of the pre-heating, no further mini-haloes can form in the entire comoving volume reionized by this mini-halo. As a result, the first generation of stars is expected to form as a population of single, isolated stars. Star formation ensues again only once more massive haloes with deeper potential wells aggregate. These rare, high-density peaks are likely to coincide with the highly biased regions where the first isolated stars had already formed. At the time of formation of deeper potential wells, such sites were inevitably already polluted with metals, unless the first stars collapsed directly to black holes without associated metal production. As a result, the transition from Population III to II (metal-free to normal) stellar populations is associated with the halo mass scale. This is in contrast to scenarios (Cen 2003; Wyithe & Loeb 2003) where star formation can continually proceed in lower-density peaks, which are far from the initial sites of star formation and still contain relatively pristine gas (allowing metal-free star formation to continue to relatively low redshifts). Given the factor of $\sim 10$–20 difference in the ionizing photon production efficiency per unit mass between Population III and II stars (Tumlinson & Shull 2000; Bromm, Kurkutzi & Loeb 2001b; Schaerer 2002), this will have important consequences for the redshift evolution and topology of reionization.

(v) 21-cm observations. It has been suggested that mini-haloes will be observable at the redshifted 21-cm line frequency both in emission (Iliev et al. 2002a,b), and absorption (Furlanetto & Loeb 2002). If mini-halo formation is strongly suppressed, these two windows on small-scale structure during the cosmological Dark Ages will disappear. 21-cm observations of mini-haloes therefore provide an interesting probe of the topology of fossil H II regions: if mini-haloes are seen, that comoving patch of the IGM has never been ionized. This is complementary to other observations such as Gunn–Peterson absorption, which only probe the instantaneous ionization state.

The rest of this paper is organized as follows. In Section 2, we compute the `entropy floor’ due to early mini-haloes. In Section 3, we calculate the effects of this entropy floor on the gas density profile in mini-haloes. In Section 4, we discuss and quantify several important effects of this entropy floor for reionization, and in Section 5 we explicitly calculate the effect on the reionization history. We summarize our findings and discuss their implications in Section 6. Throughout this paper, we adopt the background cosmological parameters as measured by the WMAP experiment (Spiegel et al. 2003, tables 1 and 2), $\Omega_m = 0.29$, $\Omega_\Lambda = 0.71$, $\Omega_b = 0.047$, $h = 0.72$ and an initial matter power spectrum $P(k) \propto k^n$ with $n = 0.99$ and normalization $\sigma_8 = 0.9$.

2 ENTRPY FLOOR DUE TO EARLY REIONIZATION

2.1 Entropy floor in Fossil H II regions

Early reionization introduces an ‘entropy floor’ in the intergalactic medium that impedes gas accretion and cooling in haloes with
The spherical overdensity $\delta$ is defined as $\delta = \rho_{\text{halo}} / \rho_{\text{background}} - 1$, where $\rho_{\text{halo}}$ is the density of the halo and $\rho_{\text{background}}$ is the mean density of the universe. When $\delta$ is small, the entropy is conserved. In the low-density IGM, Compton cooling is the dominant source of gas cooling, while in high-density halos, atomic cooling becomes dominant. The parameter dependence of the entropy floor is thus

$$K = \frac{T}{n_{\text{H2}}^{2/3}} = 100 \left( \frac{k_B T}{1 \text{ eV}} \right) \left( \frac{n_{\text{H2}}}{10^{-3} \text{ cm}^{-3}} \right)^{-2/3} \text{ eV cm}^2 \quad (3)$$

as the 'entropy' of the gas, even though it is not strictly the thermodynamic entropy $S \propto \log (K)$. This is a useful convention because $K$ is conserved when the gas evolves adiabatically. Thus, it is conserved during Hubble expansion, and during accretion on to haloes (provided the accretion shock is weak, gas will be accreted isentropically). In Fig. 2, we show the dependence of the final entropy $K$ on redshift (the dependence on $T_{\text{c}}$, $t$ follow simply from Fig. 1). The gas has significantly lower entropy at higher redshift, since Compton cooling is more efficient. However, the final entropy at a given redshift depends only weakly on the overdensity $\delta$. Denser gas remains at higher temperature, since it recombines faster than it cools, and the combination of higher temperature and higher density roughly cancel $K \propto T \delta^{-2/3}$. We can therefore ignore the weak $\delta$ dependence of $K$, and assume that it depends only on $z$. The entropy floor is thus

$T_{\text{c}} < 10^9 \text{ K}$, which cannot excite atomic cooling. Early reionization even introduces Jeans smoothing effects in haloes with $T_{\text{c}} > 10^4 \text{ K}$, since the strength of the accretion shock is weaker in pre-heated gas, and a deeper potential well is required to collisionally ionize the gas. The overall effect is to introduce a substantial core in the density profile. The effect is similar to the presumed pre-heating of the IGM by galactic winds or AGN outflows at low redshift. There, the finite entropy of the gas introduces a core in group and cluster density profiles and is thought to be responsible for the deviation from self-similarity in the observed cluster $L_X - T_X$ relation (for a recent review, see Rosati, Borgani & Norman 2002). Entropy is more fundamental than gas temperature in low-redshift clusters and high-redshift mini-haloes: in both cases, entropy is conserved, since $t_{\text{cool}} \gg t_{\text{H}}$. We exploit this analogy, and deliberately employ language and techniques from the well-studied low-redshift case to study high-redshift mini-haloes. Here, we calculate the expected level of the 'entropy floor' due to early reionization.

Reionization is likely to be a highly stochastic process where sources ionize a patch of the IGM and then fade; the fossil H II region subsequently cools and recombines. What is the final temperature a fossil H II region can cool to? At the high redshifts of interest, Compton cooling off the CMB is by far the dominant source of gas cooling in the low-density IGM (e.g. at $z = 19$ and $T = 10^3 \text{ K}$, $t_{\text{C}} \sim 0.1 t_{\text{cool}}$, where $\delta$ is the gas overdensity and $t_{\text{cool}}$ is the atomic cooling time-scale). The final gas temperature is therefore determined by the competition between Compton cooling and hydrogen recombination: after the gas recombines, it 'decouples' from the CMB, and Compton cooling is no longer efficient. The Compton cooling time-scale is independent of density and temperature:

$$t_{\text{c}} = 3 m_e c \left( 4 x_{\sigma T} a T_{\text{CMB}} \right)^{-1} = 1.4 \times 10^7 \left( \frac{1 + z}{20} \right)^{1/4} \text{ yr}, \quad (1)$$

whereas the recombination time

$$t_{\text{rec}} = 3.9 \times 10^4 \left( \frac{1 + z}{20} \right)^{-3} \delta^{-1} \text{ yr} \quad (2)$$

is shorter for higher-density gas. Thus, denser gas 'decouples' from the CMB earlier and freezes out at higher temperatures. For the redshifts of interest, $t_{\text{rec}}(\delta = 1) > t_{\text{c}}$, allowing for substantial cooling below $\sim 10^4 \text{ K}$.

We can study the temperature evolution of a cooling and recombinating parcel of gas $T(t, T_i, \delta, z)$ as a function of time $t$ (or redshift interval $\delta z$), initial temperature $T_i$, gas overdensity $\delta$, and redshift $z$. In Fig. 2 (below) we show the dependence of the final temperature on these parameters by solving the full set of chemical evolution equations for primordial gas. We neglect H$_2$ chemistry as a trace UV flux is sufficient to photodissociate H$_2$ at these low IGM densities. Our fiducial patch has an initial temperature $T_i = 20000 \text{ K}$, cools at $z = 19$, for a time interval $t = 0.6 \delta t_{\text{H}}(z = 19) = 1.9 \times 10^3 \text{ yr}$ (corresponding to roughly $\delta z \approx 7$), and lies at an overdensity $\delta = 1$. The gas cools most effectively at high redshift (when the Compton cooling time is short) and at low overdensity (when it decouples from the CMB at late times). For $z = 19$ and $\delta = 10$, (which is perhaps characteristic of gas being accreted on to haloes) the gas remains at a few $\times 10^3 \text{ K}$. The gas can therefore ignore the weak $\delta$ dependence of $K$, and assume that it depends only on $z$. The entropy floor is thus
and the increased temperature and density roughly cancel. 

δ overdensity less ef

The gas retains more entropy at lower redshift, since Compton cooling is

overdensity δφ(33) and (35). As in Fig. 1, our fiducial patch has T_s = 1, δ = 1, z = 19, 

t = 0.5t_H(z = 19). The dependence on T_s and t are the same as in Fig. 1. 

The gas retains more entropy at lower redshift, since Compton cooling is

less efficient. However, the final entropy shows only a weak dependence on

overdensity δ: the gas cools out at higher temperatures at higher densities, 

and the increased temperature and density roughly cancel.

It is useful to define a parameter that compares the IGM entropy

to the entropy generated by gravitational shock heating alone. Let us define the quantity 

\( \dot{K} \equiv K_{\text{IGM}} / K_0 \), where K_0 = T_vir / n(r_{vir})^2/3, and n(r_{vir}) = (\Omega_b / \Omega_m)NFW(r_{vir})/(\mu m_p) [\text{and } \rho_{NFW} 

is the NFW (Navarro, Frenk & White 1997) dark matter density profile]. This is the entropy due to shock heating alone at the virial radius; the justification for this will become clearer in the ensuing section. As \( \dot{K} \) increases, the Jeans smoothing effects due to finite IGM entropy become increasingly more significant. This parameter \( \dot{K} \) will be used extensively in the following sections, and it is useful here to obtain a sense of what values of \( \dot{K} \) are expected. In Fig. 3, we plot \( \dot{K} \) for a T_vir = 9000 K halo as a function of redshift, using K_{IGM}(z) shown in Fig. 2. Since \( \dot{K} \propto T_vir^{-1} \), and a T_vir = 9000 K halo is about the most massive that would still not experienceatomic cooling, the solid line depicts a lower limit on \( \dot{K} \) for all mini-haloes. Appropriate values for \( \dot{K} \) for smaller haloes can be read off simply from the \( \dot{K} \propto T_vir^{-1} \) scaling. We see that for most haloes, \( \dot{K} > 1 \); although K_{IGM}(z) falls at high redshift, this is somewhat offset by the fact that typical potential wells are much shallower, and thus K_0 also falls rapidly (as can be seen by the dashed lines, which depict \( \dot{K} \) for 2σ and 3σ fluctuations).

Typical values for \( \dot{K} \) are illustrated further in Fig. 4, where we show the mass-weighted fraction of mini-haloes that have entropy parameters less than a given \( \dot{K} \), given by

\[
  f(< \dot{K}, z) = \frac{\int_{M_{\text{MIN}}(z)} M_{\text{MIN}}(z) dM}{\int_{M_{\text{MIN}}(z)} M_{\text{MIN}}(z) dM},
\]

where \( M(\dot{K}) \) is the mass corresponding to \( \dot{K} \) for a given K_{IGM}(z), M_{\text{MIN}}(z) is the mass corresponding to a T_vir = 10^4 K halo, and

dn/dM is the Press–Schechter mass function. Virtually all haloes have \( \dot{K} > 1 \), and the median value of \( \dot{K} \) is substantially higher. Thus, the vast majority of haloes accrete gas isentropically.

2.2 Entropy floor due to X-rays

Reionization by X-rays (Oh 2001; Venketesan et al. 2001) would produce a warm (few × 100–1000 K), weakly ionized IGM with an entropy floor similar to that in fossil H\ II regions. Such X-rays could arise from supernovae, AGN or X-ray binaries, or in more exotic
models with decaying massive sterile neutrinos (Hansen & Haiman 2003). The Universe is optically thick to all photons with energies

\[ E < E_{\text{thick}} = 1.8 \left( \frac{1 + z}{15} \right)^{1/2} x_{H_1}^{1/3} \text{keV}, \]  

(5)

where \( x_{H_1} \) is the mean neutral fraction, and we have assumed \( \sigma_v \propto v^{-3} \). Thus, all energy radiated below \( E_{\text{thick}} \) will be absorbed. The relative efficiency of UV photons and X-rays in setting an entropy floor deserves a detailed separate study; here we discuss some salient points. UV photons are an ‘energetically extravagant’ means of producing an entropy floor. Most of the energy injected by UV photons is lost to recombinations and Compton cooling at high redshift; we see in Fig. 1 that the gas typically Compton cools to \( T_{\text{floor}} \sim \text{few} \times 100 \text{K} \) at the redshifts of interest. Thus, only \( \sim 10^{-2} \) of the heat injected by UV photons is eventually utilized in setting the entropy floor; all energy expended in heating the gas above \( T_{\text{floor}} \) is ‘wasted’ (it is dumped into the CMB, where it may eventually be observable as a spectral distortion, Fixsen & Mather 2002). On the other hand, in the weakly ionized gas produced by X-rays, both the recombination time and the Compton cooling time are longer than the Hubble time. In particular,

\[ t_c = \frac{1}{n} \left( \frac{x_e}{0.1} \right) \left( \frac{1 + z}{15} \right)^{-5/2} t_H, \]  

(6)

and almost none of the entropy injected is lost to cooling. Furthermore, a larger fraction of the X-ray energy goes toward heating rather than ionization [in general, a few \( \times 0.1 \) of the energy of the hot photoelectron created by an X-ray goes toward heating; this fraction quickly rises toward unity as the medium becomes progressively more ionized (Shull & van Steenberg 1985)]. Thus, if \( \epsilon_{\text{bad}}(\text{X-ray})/\epsilon_{\text{bad}}(\text{UV}) \gg 10^{-2} \) (where \( \epsilon_{\text{bad}} \) is the comoving emissivity), X-rays could be comparable or even more effective than UV photons in setting an entropy floor. The relative efficiencies of UV and X-rays is unknown, but for supernovae could be as high as (Oh 2001)

\[ \frac{\epsilon_{\text{bad}}(\text{X-ray})}{\epsilon_{\text{bad}}(\text{UV})} \approx 0.1 \left( \frac{f_{\text{esc}}}{0.1} \right)^{-1} \left( \frac{f_X}{0.1} \right), \]  

(7)

where \( f_{\text{esc}} \) is the escape fraction of ionizing UV photons from the host halo and \( f_X \) is the fraction of supernova explosion energy that goes into soft X-rays, either through inverse Compton scattering of CMB photons by relativistic electrons (Oh 2001), or free–free emission from the hot SN remnant.

To make a quick estimate, let us (fairly conservatively) assume that \( f_X \sim 3 \) per cent of the explosion energy of a supernova goes into soft X-rays. Of this, \( f_{\text{bad}} \sim 50 \) per cent of the energy goes into heating; the rest goes into secondary ionizations and atomic excitations. A supernova releases \( E_Z \sim 0.5 \text{MeV} \) in explosion energy per metal baryon, relatively independent of metallicity (a Population III ‘hypernova’ produces \( \sim 100 \) times more energy, but also \( \sim 100 \) times more metals than a standard type II SN). X-ray heating thus results in an entropy floor \( K_{\text{KM}} \approx (f_X E_Z f_{\text{bad}} Z \bar{\epsilon}) / n_{\Omega}^{2/3} \), or

\[ K_{\text{KM}} \approx 20 \text{eV cm}^{-2} \left( \frac{f_X}{0.03} \right) \left( \frac{f_{\text{bad}}}{0.5} \right) \left( \frac{Z}{10^{-3} Z_{\odot}} \right) \left( \frac{1 + z}{15} \right)^{-2}, \]  

(8)

where \( Z \) is the mean metallicity of the Universe. This is comparable to the entropy of fossil H II regions, but has a much larger filling factor, of the order of unity. A metallicity of \( Z \sim 10^{-3} Z_{\odot} \) corresponds to roughly \( \sim 1 \) ionizing photon per baryon in the Universe, both for Population II and III stars. As a result of recombinations, the filling factor of ionized regions will of course be considerably less than unity. However, the filling factor of fossil H II regions, which is the relevant quantity, will also be less than unity by a factor of \( f_{\text{overlap}} \), which represents the overlap of H II regions with fossil H II regions that have recombined. Energy spent in reionizing fossil H II regions is ‘wasted’ in terms of establishing an entropy floor. This factor is likely to be large because early galaxy formation is highly biased, and higher-mass haloes (which can resist feedback effects) will be born in regions already pre-reionized by earlier generations of mini-haloes. A fossil H II filling factor of the order of unity with little overlap can only be achieved if the formation of mini-haloes is highly synchronized (see the discussion in Section 5).

The mean free path of X-ray photons generally exceeds the mean separation between sources, becoming comparable to the Hubble length for \( \sim 2 \text{keV} \) photons. The entire Universe is thus exposed to a fairly uniform X-ray background, and the entire IGM acquires a uniform entropy floor, with an amplitude that scales with the amount of star formation, as in equation (8). Even regions far from sites of star formations, which have never been engulfed in an H II region, will be affected. In contrast, in pure UV reionization scenarios, there are large spatial fluctuations in entropy, which depend on the topology of reionization, and the redshift at which a comoving patch was last ionized.

Because of the relative uncertainty of the amplitude of the X-ray background, we use the entropy floor associated with fossil H II regions in the rest of this paper. The mass fraction of affected mini-haloes scales with the filling factor of fossil H II regions. This is therefore a minimal estimate; the filling factor could approach unity if X-rays are important.

We now consider the effects of a finite entropy floor on mini-halo gas density profiles.

### 3 GAS DENSITY PROFILES

Once Compton cooling and radiative cooling become inefficient, the gas evolves adiabatically. We can therefore compute static equilibrium density profiles, and see how they change as a function of the entropy floor. The models we construct are in the spirit of Voit et al. (2002), Tozzi & Norman (2001), Oh & Benson (2003) and Babul et al. (2002), which match observations of low-redshift cluster X-ray profiles well.

Naively, one might assume that only gas that remains at temperatures comparable to the virial temperatures of mini-haloes would suffer appreciable Jeans smoothing. Since in many cases the IGM can cool down to a few \( \times 100 \text{K} \), one might assume that this level of pre-heating would have negligible effects on the density profile of gas in mini-haloes. This is false: the important quantity is not the temperature but the entropy of the gas. Since gas in the IGM is heated at low density, it has comparatively high entropy. Gas at mean density that is heated to temperatures

\[ T_{\text{IGM}} > 90 \left( \frac{T_{\text{vir}}}{3000 \text{K}} \right) \left( \frac{\delta}{200} \right)^{-2/3} \text{K} \]  

(9)

(where \( \delta \) is the overdensity of the gas in the mini-halo in the absence of pre-heating) will have an entropy in excess of that acquired by gravitational shock heating alone. Its temperature will therefore exceed the virial temperature after infall and adiabatic compression. As the level of pre-heating increases, gas at progressively larger radii in the halo undergoes Jeans smoothing effects. We now calculate this in detail.

We first construct the default entropy profile of the gas without pre-heating. We assume that in the absence of heating or cooling

processes, the gas distribution traces that of the dark matter, an Ansatz which is indeed observed in numerical simulations (e.g. Frenk et al. 1999) (the dark matter is assumed to follow the NFW, Navarro et al. 1997, profile). This assumption becomes inaccurate at the very centre of the halo, where finite gas pressure causes the gas distribution to be more flattened and less cuspy than the dark matter density distribution. In particular, even in the absence of pre-heating the IGM has a finite temperature after decoupling from the CMB and cooling adiabatically: \( T_{\text{vir}}(z) \approx 2.73(1 + z_d)(1 + z)/(1 + z_d)^2 \), where the matter–radiation decoupling redshift \( z_d \approx 150 \). This gives rise to a finite entropy floor:

\[
K_{\text{min}} = 4.6 \times 10^{-2} \text{eV cm}^{-2},
\]

infall velocity does not exceed the local speed of sound, then the \( \sim \) are dealing with the case where the ionization fraction is small and gas traces the dark matter, hydrostatic equilibrium gives the entropy even in the absence of pre-heating. In regions where \( \text{gas is shocked to the entropy} \)

\[
K_{\text{shock}}(r) = \frac{1}{\rho g(r) r^2} \left[ \int_{r_{\text{sh}}}^r - \frac{GM(r)}{r^2} dr + \frac{\rho g(r_{\text{sh}})}{\mu m_n} K_{\text{shock}}(r_{\text{sh}}) \right].
\]

The final entropy profile is therefore \( K(r) = \max[K_{\text{min}}, K_{\text{shock}}(r)] \). (10)

Note that the mean molecular weight is \( \mu = 0.59 \) for fully ionized gas and \( \mu = 1.22 \) for fully neutral primordial gas; in this paper we are dealing with the case where the ionization fraction is small and therefore use the latter figure. This results in virial temperatures that are higher by a factor of \( \sim 2 \).

What is the effect of pre-heating on the entropy profile? If the infall velocity does not exceed the local speed of sound, then the gas is accreted adiabatically and no shock occurs; the gas entropy \( K_{\text{IGM}} \) is therefore conserved. If gas infall is supersonic, then the gas is shocked to the entropy \( K_{\text{shock}} \) computed in equation (11). Tozzi & Norman (2001) found that the transition between the adiabatic accretion and shock heating regime is very sharp. To a very good approximation \( K(M) = \max(K_{\text{shock}}(M), K_{\text{IGM}}) \); with this new entropy profile we can compute density and temperature profiles. In fact, for most high-redshift mini-haloes the 'pre-heating' entropy exceeds the shock entropy even at the virial radius, so the gas mass is accreted isentropically. This occurs when the entropy floor exceeds

\[
K_{\text{IGM}} > 2.3 \text{eV cm}^{-3} \left( \frac{T_{\text{vir}}}{5000 \text{K}} \right)^{-2/3} \left( \frac{\delta}{50} \right)^{-2} \left( \frac{1 + z}{20} \right)^{-2}, \quad (12)
\]

where the gas overdensity \( \delta \approx 50 \) at the virial radius in the absence of pre-heating is a weak function of the NFW concentration parameter \( c \). Although \( c \) depends weakly on the collapse redshift, for simplicity we shall assume in this paper that \( c = 5 \) for all mini-haloes.

We have compared our entropy profiles calculated using equation (11) with the prescription in Tozzi & Norman (2001) (and also adopted by Oh & Benson 2003). In this method an accretion history for a halo is prescribed via extended Press–Schechter theory; this allows one to compute the strength of the accretion shock and thus the gas entropy (using standard Rankine–Hugoniot jump conditions) for each Lagrangian mass shell. The two methods agree extremely well; we therefore use equation (11) for both speed and simplicity.

Given an entropy profile, from hydrostatic equilibrium the density profile of the gas is then given by

\[
\rho(r) = \tilde{K}(r)^{-1/\gamma} P(r_{\text{sh}})^{-1/\gamma} + \frac{\gamma - 1}{\gamma} \int_{r_{\text{sh}}}^r - dr \frac{GM(r)}{r^2} \tilde{K}(r)^{-1/\gamma} r^{-1/(\gamma - 1)},
\]

where \( \tilde{K} \equiv P/\rho^{\gamma/\gamma} = k_B(\mu m_n)^{-3/3} K \). The temperature profile can then be determined from \( T = (\mu m_n/k_B) \tilde{K}^{1/4} \).

The solution of this equation requires a boundary condition which sets the overall normalization, here expressed as \( P(r_{\text{sh}}) \). There is some ambiguity in this choice. The often used boundary condition \( M_f = f_B M_{\text{halo}} \) at \( r = r_{\text{sh}} \) is unphysical as it does not take into account the suppression of accretion due to finite gas entropy. We make the following choice. Let us define \( P_{\text{shock}}(r_{\text{sh}}) \equiv (\Omega_m/\Omega_{\text{nu}}) \rho(r_{\text{sh}})/(\mu m_n)K_{\text{shock}}(r_{\text{sh}}) \); the pressure at the virial radius due to shock heating alone. For \( K_{\text{shock}}(r_{\text{sh}}) > K_{\text{IGM}} \), the final conditions at the virial radius are not strongly affected by the entropy floor, since the shock boosts the gas on to a new adiabat. We therefore set \( P(r_{\text{sh}}) = P_{\text{shock}}(r_{\text{sh}}) \). The boundary condition must change when \( K_{\text{IGM}} > K_{\text{shock}}(r_{\text{sh}}) \), when accretion takes place isentropically, and the entropy floor is fundamental in determining the gas pressure. In this case, \( P(r_{\text{sh}}) \) (which is essentially a constant of integration) is chosen so that \( \rho(r) \to \rho_s \) as \( r \to \infty \). The latter boundary condition implicitly assumes that hydrostatic equilibrium prevails beyond the virial radius. This is questionable, but is arguably reasonable:

\[
L_{\text{sh}} \approx c_{\text{sh}} h^{-1} \left( \frac{K_{\text{IGM}}}{3 \text{eV cm}^{-2}} \right)^{1/2} \left( \frac{1 + z}{20} \right)^{-1/2} \left( \frac{\delta}{10} \right)^{1/3},
\]

which is approximately five times larger than the virial radius for a \( \sim 6000-\text{K} \) halo at the same redshift. Indeed, by definition \( L_{\text{sh}} > r_{\text{vir}} \) for Jeans smoothing effects to be important.

Having established the appropriate boundary conditions, we can now compute detailed gas profiles. The entropy, density, pressure and temperature profiles as a function of \( \tilde{r} \equiv r/r_{\text{sh}} \) are shown in Fig. 5. These profiles are of course universal and independent of the halo mass once \( \tilde{K} \) is set, if the weak dependence of the NFW concentration parameter \( c \) with mass is ignored. As the entropy floor increases, the central pressure and density decline, while the central.

Figure 5. The dimensionless entropy \( \tilde{K} = K/K_0 \), pressure \( \tilde{P} = P/P_0 \), temperature \( \tilde{T} = T/T_0 \), and density \( \tilde{\rho} = \rho/\rho_0 \), as a function of radius \( \tilde{r} = r/r_{\text{sh}} \). Here, \( K_0, P_0, T_0 \) are the values of these quantities at \( r_{\text{sh}} \), without pre-heating, while \( \rho_0 = \rho_{\text{sh}} \), the mean baryonic density. The entropy profile \( \tilde{K}(r) \) uniquely specifies \( \tilde{P}(\tilde{r}), \tilde{T}(\tilde{r}), \tilde{\rho}(\tilde{r}) \), independent of halo mass or redshift.
which the gas mass fraction $f_g$ is continuous at $\hat{K} = 1$, when we switch from one boundary condition to another. This need not have been the case, and gives us confidence that we handle the transition to isentropic accretion correctly. We see that realistic levels of the entropy floor (as computed in Section 2) cause a substantial depression in gas fractions in mini-haloes.

It is interesting to compare our derived gas fractions with other estimates. For the case where $K_{\text{IGM}} > K_{\text{acc}}(r_{\text{vir}})$, accretion takes place isentropically. The halo will therefore accrete gas at roughly the adiabatic Bondi accretion rate (e.g. Balogh, Babul & Patton 1999):

$$M_b \approx 1.86\pi\lambda G^2 M_{\text{halo}}^2 K_{\text{IGM}}^{-3/2}.$$  \hspace{1cm} (15)

The total accreted gas mass is then roughly $M_{\text{gas}} \approx \min f M_{\text{vir}}(\Omega_\rho/\Omega_m) M_{\text{halo}}$, where $f$ is some unknown normalization factor, which takes into account the fact that the total halo mass is not constant but was lower in the past (and hence that the gas accretion rate was lower in the past). Another estimate which is a good fit to the results of hydrodynamic simulations is (Gnedin 2000)

$$M_b \approx \frac{(\Omega_\rho/\Omega_m) M_{\text{halo}}}{1 + (2^{1/3} - 1)M_{\text{halo}}^2/M_{\text{halo}}}.$$  \hspace{1cm} (16)

There is only one free parameter: $M_{1/2}$, the mass of the halo in which the gas mass fraction $f_g = 0.5$. Gnedin (2000) shows that $M_{1/2}$ is well approximated by the ‘filtering mass’ $M_F$. However, $M_F$ depends on the unknown thermal history of the IGM. To make a self-consistent comparison, we compute $M_{1/2}$ with the density profiles computed using our fiducial boundary conditions. Interestingly, we find that $M_{1/2}$ roughly corresponds to the halo mass when accretion begins to take place isentropically $\hat{K} \sim 1$.

The results are shown in Fig. 6. All three estimates agree well (note that the normalization $f$ of the Bondi accretion prediction is a free parameter; the plot shown is for $f = 0.3$). The widely used Gnedin (2000) fitting formula predicts even lower gas fractions (and as we will see, clumping factors) at high entropy levels $\hat{K} \gg 1$. On the other hand, the slope of the $f_\Omega(\hat{K})$ relation for our boundary conditions agrees very well with the Bondi accretion prediction in this regime. Our boundary conditions therefore yield fairly conservative estimates of the effects of pre-heating. Moreover, unlike these other estimates, we are able to compute detailed density profiles, which is crucial for some of our later calculations.

There are two other quantities which are of particular interest when computing density profiles. One is the central density $\rho_c$, which affects the ability of mini-haloes to form H$_2$ in the face of reionization. Another is the gas clumping of the halo, defined as

$$C_{\text{halo}} = \langle \rho^2 \rangle / \langle \rho \rangle^2.$$  \hspace{1cm} (17)

where the angled brackets indicate a volume-averaged quantity,

$$\langle X \rangle = \frac{1}{V} \int_0^{r_{\text{vir}}} dr 4\pi r^2 X.$$  \hspace{1cm} (18)

Note that $C_{\text{halo}} \geq 1$ always. The clumping factor $C_{\text{halo}}$ plays a central role in determining the photon budget required for reionization; we evaluate the global clumping factor in the next section. In Fig. 7, we show the effect of increasing the entropy parameter $\hat{K}$ on the central density and the clumping factor. Both decline rapidly with $\hat{K}$. We shall use these two results in the following sections.

Figure 6. The gas fraction within the virial radius $f_g = (M_g/M_{\text{halo}})/(\Omega_\rho/\Omega_m)$ as a function of the entropy parameter $\hat{K}$. The solid line indicates our fiducial boundary conditions; the dotted line corresponds to the Gnedin (2000) fit to numerical simulations; while the dashed line corresponds to the Bondi accretion prediction (valid only for isentropic accretion, $\hat{K} > 1$). All three show good agreement. The self-similarity of the problem implies that the computed gas fractions apply to mini-haloes of all virial temperatures, provided $\hat{K}$ is appropriately scaled. For illustrative purposes, appropriate values of $\hat{K}$ for a mini-halo of $T_{\text{vir}} = 5000$ K at $z = 20, 13$ are shown (see Fig. 3).

Figure 7. The effect of increasing the entropy parameter $\hat{K}$ on the central gas density $\rho_c$, and gas clumping $C_{\text{halo}}$, top panel, and as defined in equation (17). The solid lines correspond to our fiducial boundary conditions, while the dashed lines correspond to adjusting $P(r_{\text{vir}})$ to reproduce the gas fractions from the Gnedin (2000) fit to numerical simulations (the results for the Bondi accretion prediction are almost identical to the solid curve). Both $\rho_c$ and $C_{\text{halo}}$ decline rapidly as $\hat{K}$ increases.
We now use these gas density profiles to compute global effects of mini-halo suppression.

4 Global Effects of Mini-Halo Suppression

4.1 Suppression of collapsed gas fraction and gas clumping

An entropy floor suppresses the fraction of gas which is bound within mini-haloes. As we discuss in Section 4.4, the mean 21-cm emission from mini-haloes is directly proportional to this global gas fraction; if it is strongly suppressed the signal will be unobservable. The global collapsed gas fraction is

\[ f_{\text{gas}} = \frac{1}{\rho_b} \int_{M_{l}}^{M_i} \frac{dM}{dM} \frac{d \Omega}{d \Omega} f_{\text{halo},g}(\hat{K}) M, \]

where \( f_{\text{halo},g} \) is the halo gas fraction as in Fig. 6, and \( \rho_b \) is the comoving baryon density. This is shown in the bottom panel of Fig. 9 (below). Curves are shown for no pre-heating \( K_{\text{IGM}} = K_{\text{min}} \), fixed values of the entropy floor \( K_{\text{IGM}} = 1, 10 \) eV cm\(^2\) (which correspond to the IGM settling to some temperature at a given redshift and remaining at constant entropy thereafter), and the redshift-dependent entropy \( K_{\text{IGM}}(z) \) shown in the top panel of Fig. 2, which reflects the increasing entropy of gas that Compton cools at late epochs. For realistic values of \( K_{\text{IGM}} \), the collapsed gas fraction is suppressed by one to two orders of magnitude.

It is interesting to plot the effect of pre-heating on the baryonic mass function. This can be computed simply as

\[ \frac{dN}{dM} = \frac{dN}{dM_b} \frac{dM_b}{dM}. \]

We show this in Fig. 8, at \( z = 10, 15, 20 \) for \( K_{\text{IGM}}(z) \) as in the top panel of Fig. 2, and for a fixed entropy floor \( K_{\text{IGM}} = 3 \) eV cm\(^2\). As expected, the effect of pre-heating is most drastic at low masses. Note that the baryonic mass function is very similar at all redshifts for \( K_{\text{IGM}}(z) \); this can also be seen in the bottom panel of Fig. 9, where the collapsed fraction in mini-haloes does not change appreciably with redshift.

Another quantity of great interest is the global gas clumping factor \( C = (n^2)/(n)^2 \). Gas clumping shortens the recombination time \( t_{\text{rec}} \approx 1/(\alpha n C) \) and thus increases the total number of photons per baryon required to achieve reionization. It is given by

\[ C = (1 - f_V)C_{\text{halo}} + \int_{M_{\text{h}}}^{M_{\text{i}}} \left( \frac{t_{\text{evap}}}{t_{\text{rec}}} \right) C_{\text{halo}}(\hat{K}) \frac{d \Omega}{d \Omega} V_{\text{halo}} \frac{dM}{dM}, \]

where \( f_V \) is the collapsed fraction by volume of mini-haloes, \( C_{\text{halo}} \approx 1 \) is the clumping factor of the IGM, \( C_{\text{halo}}(\hat{K}) \) is the halo clumping factor as calculated in the previous section and shown in the top panel of Fig. 7, and \( V_{\text{halo}} = (4/3) \pi r_{\text{vir}}^3 \) is the volume of a mini-halo. The factor \( t_{\text{evap}}/t_{\text{rec}} \) describes some explanation. Mini-haloes only contribute to recombinations when they are photoionized. However, once they are exposed to ionizing radiation, they will be evaporated, essentially in the sound crossing time of photoionized gas (Shapiro, Raga & Mellemma 1998; Barkana & Loeb 1999). As in Haiman et al. (2001), we therefore set the evaporation time \( t_{\text{evap}} = r_{\text{vir}}/(10 \) km \( s^{-1}) \), and weight the halo clumping factor by the duty cycle \( t_{\text{evap}}/t_{\text{rec}} \).

The clumping factor in equation (21) is a lower limit to the total gas clumping; there will be an additional contribution from larger scales as well. However, the clumping due to mini-haloes is dominant at high redshifts of interest.

Our result is shown in the top panel of Fig. 9. For the no entropy floor \( K_{\text{IGM}} = K_{\text{min}} \) case (top curve), gas clumping is much lower at high redshift, since fewer mini-haloes have collapsed. Once a patch of IGM is reionized early at high redshift, then an entropy floor is established and subsequent gas clumping is suppressed. Thus, early star formation reduces the total photon budget required to achieve full reionization.

\[ \text{Figure 8. The mass function of baryons, as in equation (20), for } z = 10, 15, 20. \text{ Dark solid curves show the mass function for } K_{\text{IGM}} = K_{\text{min}} \text{, light solid curves are for redshift-dependent entropy } K_{\text{IGM}}(z) \text{ as in the top panel of Fig. 2, and dashed curves are for a fixed entropy floor } K_{\text{IGM}} = 3 \text{ eV cm}^2. \text{ Note that the baryonic mass function is very similar at all redshifts for } K_{\text{IGM}}(z). \text{ For } T_{\text{evap}} > 10^4 \text{ K, an entropy floor has no effect on the mass function, which reverts back to the dark solid curves (we do not extend the light solid and dashed curves above the corresponding halo masses).} \]

\[ \text{Figure 9. The evolution of the collapsed gas fraction in mini-haloes (lower panel) and the gas clumping factor (upper panel) with redshift. The results are shown for no pre-heating (with } K_{\text{IGM}} = K_{\text{min}} \text{, and constant values of entropy } K_{\text{IGM}} = 1, 10 \text{ eV cm}^2\text{ and the redshift-dependent entropy } K_{\text{IGM}}(z) \text{ shown in the top panel of Fig. 2. Both the global gas fraction in mini-haloes and gas clumping are strongly suppressed for realistic values of the entropy floor. In particular, } C \rightarrow 1 \text{ for realistic values of } K_{\text{IGM}}, \text{ greatly reducing the photon budget required for reionization.} \]
an order of magnitude even for the low entropy levels $\sim 1$ eV cm$^{-2}$ associated with reionization at high redshift $z \sim 20$. For the redshift-dependent entropy $K_{\text{HM}}(z)$ shown in the top panel of Fig. 3, mini-halo clumping is strongly suppressed, and $C \approx 1$.

4.2 Suppression of H$_2$ formation

The formation of H$_2$ molecules at high redshift have long been thought to be critical to gas cooling and the earliest episodes of star formation (Couchman & Rees 1986). It has also long been shown that H$_2$ molecules are fragile, and subject to photodissociative feedback from an early UV background in the Lyman–Werner bands of H$_2$ (Haiman, Rees & Loeb 1997). This negative UV feedback on the H$_2$ abundance has since been studied by numerous authors (e.g. Ciardi et al. 2000; Haiman et al. 2000; Machacek et al. 2001; Ricotti et al. 2001, 2002a,b).

An important point is that H$_2$ formation and cooling in minihalos will be further suppressed relative to the no-pre-heating case. This is because the finite entropy of the gas allows it to resist compression; thus, the gas is at considerably lower density. Collisional processes with cooling times $t_{\text{cool}} \propto 1/n$ will be suppressed relative to radiative processes that photodissociate H$_2$. It has been recognized that gas profiles should have a central core rather than a cusp (e.g. Shapiro, Iliev & Raga 1999) and that lower central densities will affect H$_2$ chemistry (Haiman, Rees & Loeb 1996; Tegmark et al. 1997). However, the fact that pre-heating can produce a much larger, lower-density core than hitherto considered, and the subsequent implications for H$_2$ formation and cooling, has not been explored.

It has been shown that H$_2$ formation is enhanced in fossil H II regions due to increased electron fraction there (Ricotti et al. 2001). Although H$_2$ formation is very slow in the low-density IGM, a high initial electron abundance will persist when weakly ionized gas is accreted on to non-linear structures, and aid H$_2$ formation at that point (Oh 2000). The relevant question is whether the gas can reach sufficiently high densities so that the two-body H$_2$ formation and cooling processes (catalysed by the enhanced electron fraction) dominate over H$_2$ photodissociation. Ricotti et al. (2002b) have indeed found that a net positive feedback occurs only inside fluid elements that are already dense when they are photoionized. Here we start from gas that is near the IGM density, and ask whether this gas can achieve sufficiently high densities (by contracting on small scales, below the resolution of cosmological simulations) for a similar net positive effect on the H$_2$ abundance. We can conservatively take into account the H$_2$ enhancement by assuming the maximum initial abundance of H$_2$, $x_{\text{H}_2} \sim 10^{-3}$; this hard upper limit is independent of density or ionization fraction and is due to ‘freeze-out’ (Oh & Haiman 2002 hereafter OH02). The H$_2$ abundance can only exceed this if three-body processes are important, which only takes place at very high density $n > 10^4$ cm$^{-3}$.

In OH02, we showed that the minimum temperature $T_{\text{min}}$ a parcel of gas can cool to depends almost exclusively on $t_{\text{cool}}/t_{\text{diff}} \propto J_{\text{UV}}/n$ (see fig. 6 of OH02). This scaling behaviour is only broken when the gas approaches high densities $n > 10^4$ cm$^{-3}$ (at this point, the cooling time becomes independent of density and $t_{\text{cool}}/t_{\text{diff}} \propto J_{\text{UV}}$). However, the latter regime is never reached in pre-heated gas, which is at much lower densities ($n \sim 10^3$ cm$^{-3}$ corresponds to $\delta \sim 6 \times 10^8$ at $z = 19$). In subsequent discussion, we shall assume the dependence of $T_{\text{min}}$ on $J_{\text{UV}}/n$ shown in fig. 6 of OH02.

Thus, in the presence of a radiation field $J_{\text{UV}}$, the gas must be at a minimum density $n_{\text{min}}$ to cool down to a temperature $T_{\text{min}}$. We have already calculated the maximum central density $n_c$ of gas in a halo given an entropy parameter $\hat{K}$, as in Fig. 7. If the central density is less than the critical density, $n_c < n_{\text{crit}}$, then none of the gas in the halo can cool down to $T_{\text{min}}$. For a given entropy parameter $\hat{K}$, there is therefore a minimum radiation field $J_{\text{UV}}$ above which no gas can cool down to $T_{\text{min}}$. We plot this in the top panel of Fig. 10. This plot is valid for all haloes at all redshifts provided $\hat{K}$ and $J_{\text{UV}}$ are both rescaled appropriately [note that since $n \propto (1 + z)^3$, to keep $J_{\text{UV}}/n$ constant, $J_{\text{UV}} \propto (1 + z)^3$]. To get a sense of typical values of $J_{\text{UV}}$, the radiation field corresponding to $n_c$ ionizing photons per baryon in the Universe is $J_{\text{UV}} \approx (h_p/4\pi)n_e n_B(1 + z)^3/10^{21}$ erg s$^{-1}$ cm$^{-2}$ Hz$^{-1}$ sr$^{-1}$ $\approx 10n_r[(1 + z)/16]^{1/3}$ (where $n_r$ is the comoving baryon number density and $h_p$ is the Planck constant).

We see that an entropy floor greatly reduces the radiation field required to prevent gas cooling. For reasonable values of $\hat{K}$, the reduction can be as much as four orders of magnitude. Thus, even if there is only a very weak radiation field, in the presence of an entropy floor effective H$_2$ cooling and star formation will be quenched. Note also that the cooling time exceeds the Hubble time for large values of $\hat{K}$. This regime is reached only when gas densities are so low that even a very weak radiation field can dissociate H$_2$: $t_{\text{cool}} \sim t_{\text{diff}} \sim 2 \times 10^8$ (J$_{\text{UV}}/10^{-4}$)$^{-1}$ yr.

Without an entropy floor, the density profiles of mini-haloes at a given redshift are self-similar; thus, roughly comparable fractions of their gas can cool down to $T_{\text{min}}$ and will be available as fuel for star formation. In the presence of an entropy floor, the fraction of gas with $n > n_{\text{gas}}$ is greatly reduced, and much less gas can cool. Moreover, the self-similarity is broken: shallower potential wells (which have higher $\hat{K}$) are more strongly affected by an entropy floor, and their central gas densities are much lower. Using our halo density profiles, we can calculate the fraction of gas above the critical density, $f_{\text{gas}} = M_b(>n_{\text{crit}})/[\Omega_{\text{DM}}/\Omega_{\text{gas}}]M_{\text{halo}}$. We plot this in the lower panel of Fig. 10. At some critical value of $\hat{K}$, the fraction of gas that can cool plummets dramatically. This value of $\hat{K}$ corresponds roughly to when the low-density core falls below the critical density.
One possible caveat is that since fossils are likely to be metal-enriched, gas cooling is dominated by metal line cooling, rather than H$_2$ cooling. We argue that this is unlikely, for two reasons. (i) Metal line cooling only dominates H$_2$ cooling above a critical metallicity $Z > 10^{-2}Z_{\odot}$ (Hellsten & Lin 1997; Bromm et al. 2001a). This level of metal enrichment corresponds to $\sim 1 f_{UV}^{-1}$ ionizing photon per baryon in the Universe (where $f_{x}$ is the volume filling factor of metals). Thus, metal line cooling becomes significant only at late times. (ii) After an ionizing source turns off and explodes as a supernova, the metal-polluted region is much smaller than the fossil H~II region (Madau, Ferrara & Rees 2001). Most of the volume is therefore still of pristine composition, and undergoes the entropy floor suppression we have described. The metal-polluted region lies close to the high-density peak where the very first stars formed, where in any case, more massive haloes $T_{vir} > 10^7$ K will collapse. To summarize: metal line cooling is therefore unlikely to spoil our assumption of adiabaticity in most of the volume of the fossil H~II region. The exception is at the most highly biased density peaks, where metals can be thought of as effectively increasing the star formation efficiency.

The net result is that entropy injection greatly boosts the negative feedback from early star formation: the entropy floor in reionized regions results in low-density cores in the centre of haloes, in which H$_2$ is easily photodissociated by a weak external UV radiation field.

### 4.3 Negative feedback from X-rays

It is often argued that X-rays boost H$_2$ production and cooling in mini-haloes, by penetrating deep into the dense core and increasing the free-electron fraction, which is critical for gas phase H$_2$ production. Thus, X-rays are thought to exert a positive feedback effect, counter-acting photodissociation by UV radiation (Haiman et al. 1996, 2000). We show here by fairly general arguments that X-rays, in fact, exert a strong negative feedback effect, due to the entropy that they inject into the IGM. This prevents gas from compressing to sufficiently high density to produce H$_2$ and cool efficiently, and far outweighs any positive feedback from increasing the free-electron fraction.

For the sake of definiteness, we consider an intrinsic source spectrum of the form

$$J = J_{UV} \left( \frac{\nu}{\nu_c} \right)^{-\alpha} \text{h} \nu < 13.6 \text{eV},$$

$$= J_X \left( \frac{\nu}{\nu_c} \right)^{-1} \text{13.6 eV} < \nu < 10 \text{keV},$$

(22)

where $h\nu_c = 13.6$ eV and $J_{UV}/J_X = f_{break} = 1 \sim 100$ is the spectral break at the hydrogen Lyman edge; note that $f_{break} \geq 1$ always. A spectrum of the form $J_x \propto \nu^{-\alpha}$ is characteristic of the mean spectra of quasars (Elvis et al. 1994), as well as inverse Compton emission from high-redshift SNRs (Oh 2001). It is distinctive in that $V\nu_{c}$ = constant, i.e. there is equal power per logarithmic interval (though our arguments can be generalized to other spectra). Note that the actual ionizing spectrum at any given point in the IGM will be much harder, since photoelectric absorption hardens the spectrum away from the source.

What is the heating associated with this radiation field? As we argued in Section 2.2 above, all energy in the X-ray radiation field with $E_{\text{thick}} < E < E_{\text{thick}}$ (as defined in Section 5) will be absorbed by the IGM, since the Universe is optically thick at these frequencies. The radiation field may be subdivided into two components. The ‘mean field’ consists of photons with $E > E_{\text{overlap}}$, where $E_{\text{overlap}}$ is...
defined by
\[ \lambda_{\text{diff}}(E) > n^{-1/3}_{\text{source}} \quad \text{for} \quad E > E_{\text{overlap}} \] (23)
i.e, photons with \( E > E_{\text{overlap}} \) (typically, \( E_{\text{overlap}} \sim 100 \) eV) have a mean free path greater than the mean separation between sources, so that a homogeneous ionizing background is established. The ‘fluctuating field’ consists of photons with \( E < E_{\text{overlap}} \); this component is dominated by radiation from a single source, and is subject to large Poisson fluctuations. We are interested in the heating due to the ‘mean field’, which can lead to an entropy floor even outside the \( \text{H}_\text{II} \) regions of ionizing sources. It is
\[ E_{\text{heat}} = f_{\text{heat}} \int_{v_{\text{overlap}}}^{v_{\text{thick}}} \text{d}v e_v \approx f_{\text{heat}} \frac{4\pi}{l_H} J_X v_L \ln \left( \frac{v_{\text{thick}}}{v_{\text{thin}}} \right), \] (24)
where \( f_{\text{heat}} \) is the fraction of energy of the hot photoelectron created by an X-ray which goes into heating. We have conservatively set \( e_v \approx (4\pi J_X)/l_H \); this underestimates \( e_v \) for a given \( J_X \). For the spectrum we have chosen, the result is only logarithmically sensitive to the integration bounds; we set \( \ln\left(v_{\text{thick}}/v_{\text{thin}}\right) \approx 2.3 \). For steeper spectra, \( J_X \propto v^{\alpha} \) where \( \alpha > 1 \), the lower integration limit becomes important.

Over a Hubble time \( t_{15} \), the internal energy density \( U = nk_B T \approx E_{\text{heat}}/n_H \) of the gas becomes
\[ U \approx f_{\text{heat}} \frac{4\pi}{c} J_X v_L \ln \left( \frac{v_{\text{thick}}}{v_{\text{thin}}} \right). \] (25)
This make sense: since some fraction \( f_{\text{heat}} \) of the radiation field goes directly into heat, \( f_{\text{heat}}(4\pi J_X)/c \approx U_c \). Thus, X-rays will heat the IGM to a temperature
\[ T = 170 \left( \frac{f_{\text{heat}}}{0.5} \right) \left( \frac{J_X}{10^{-2}} \right) \left( \frac{1 + z}{15} \right)^{-3} \] (26).

There is therefore a direct relation between the X-ray radiation field and the entropy floor:
\[ K_{\text{IRM}} = 1.9 \text{ eV cm}^2 \left( \frac{f_{\text{heat}}}{0.5} \right) \left( \frac{J_X}{10^{-2}} \right) \left( \frac{1 + z}{15} \right)^{-5}. \] (27)

As we argued in Section 2.2 above, the IGM evolves adiabatically for \( x_e < 0.1(1 + z)/15 \)\(^{1/3}\). What level of the radiation field does this correspond to? For \( x_e \ll 0.1 \), approximately \( f_{\text{ion}} \sim 1/3 \) of the injected energy goes toward producing ionizations; the number of photoionizations is simply \( E_{\text{photo}}/(37 \text{ eV}) \). Thus, we obtain
\[ x_e = 2.3 \times 10^{-4} \left( \frac{f_{\text{ion}}}{0.3} \right) \left( \frac{J_X}{10^{-2}} \right) \] (28)
and so for \( J_X < 5 \), the IGM evolves adiabatically [note also that since \( t_{\text{rec}} \approx 10^4(1/\epsilon_{10})^{-1}(T/100 \text{ K})^{-0.7}(1 + z)/15 \)^{3/2}, we can ignore recombinations].

Given the entropy floor from equation (27), we can now calculate density profiles of haloes. In particular, we can calculate the maximum central density of a halo from the entropy floor induced by a given X-ray background. As before, from the central density of a halo we can calculate the critical level of the LW flux \( J_{\text{UV}} \) required to suppress \( \text{H}_2 \) cooling down to some temperature \( T_{\text{min}} \). This critical flux is shown as a function of \( f_X = J_X/J_{\text{UV}} \) in Fig. 12, for a halo of \( T_{\text{vir}} = 5000 \) K at \( z = 19 \). The curves may be simply rescaled for haloes of different mass at different redshifts by translating the \( x \)-axis; \( f_X \propto T_{\text{vir}}(1 + z)^3 \). The effect of X-rays increases for less massive haloes (which are less able to withstand an entropy floor) at lower redshifts (when gas densities are lower). Also shown are the maximum value of the X-ray background \( f_X \sim 1 \) for the hardest sources, and the typical minimum value of the X-ray background required for positive feedback to be effective \( f_X \approx 0.1 \) (Haiman et al. 2000). The shaded region therefore denotes the X-ray fluxes thought to produce positive feedback effects. We see for these large values of the X-ray background, X-rays cannot exert a positive feedback effect. The required X-ray background is so large that it would produce an unacceptably large entropy floor, which prevents the formation of dense regions where \( \text{H}_2 \) can withstand photodissociation. Instead, X-rays produce an overall negative feedback effect. In fact, values of \( f_X \approx 0.1-1 \) reduce by almost two orders of magnitude the value of the UV background required to photodissociate \( \text{H}_2 \) and prevent it from cooling down to low temperatures.

The only way of circumventing these difficulties is if X-rays turn on suddenly at some epoch. Haloes which have already collapsed will be unaffected by the entropy floor, but will undergo the usual positive feedback from X-rays. However, the subsequent generation of mini-haloes will be suppressed by the entropy floor.

4.4 21-cm observational signatures

At present, the only observational probes proposed for small-scale structure at high redshift such as mini-haloes are 21-cm observations with future radio telescopes such as the Square Kilometer Array (SKA)\(^2\) and the Low Frequency Array (LOFAR).\(^3\) It has been proposed that mini-haloes could be detected statistically in emission (Illiev et al. 2002a,b) and individually in absorption along the line of sight to a high-redshift radio source (Furlanetto & Loeb 2002). If an entropy floor exists, such signatures will be strongly suppressed.

21-cm signatures of mini-haloes therefore provide an interesting

---

Footnotes:
\(^2\) See http://www.nrao.edu/lofar
\(^3\) See http://www.astron.nl/lofar
indirect probe of reionization history: if mini-haloes are detected in large numbers at a given redshift, it would imply that the filling factor of fossil H I regions is still small at that epoch. In particular, a comoving patch where mini-haloes are seen has likely never been ionized before (which is a much stronger constraint than merely being neutral at the observed epoch). This could prove a very useful probe of the topology of reionization.

Mini-haloes are too faint to be seen individually in emission, and can only be detected statistically through brightness temperature fluctuations. Provided \( T_S \gg T_{\text{CMB}} \) (where \( T_S \) is the spin temperature), the 21-cm flux \( S \) is independent of the spin temperature, and depends only on the H I mass, \( S \propto M_{\text{HI}} \). Thus, \( S(\text{haloes})/S(\text{IGM}) \) is simply equal to the collapsed gas mass fraction in mini-haloes, which is always less than unity. The mini-halo signal is much smaller if an entropy floor exists, since the collapsed gas fraction declines rapidly; the bottom panel of Fig. 9 can simply be read off as \( S(\text{haloes})/S(\text{IGM}) \). Thus, the IGM dominates 21-cm emission.

Mini-haloes dominate only when: (i) the IGM spin temperature has not yet decoupled from the CMB; (ii) the IGM spin temperature has not yet decoupled from the CMB; (iii) the IGM spin temperature has not yet decoupled from the CMB; (iv) the IGM spin temperature has not yet decoupled from the CMB; (v) the IGM spin temperature has not yet decoupled from the CMB; (vi) the IGM spin temperature has not yet decoupled from the CMB; (vii) the IGM spin temperature has not yet decoupled from the CMB; (viii) the IGM spin temperature has not yet decoupled from the CMB; (ix) the IGM spin temperature has not yet decoupled from the CMB; (x) the IGM spin temperature has not yet decoupled from the CMB; (xi) the IGM spin temperature has not yet decoupled from the CMB; (xii) the IGM spin temperature has not yet decoupled from the CMB; (xiii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xv) the IGM spin temperature has not yet decoupled from the CMB; (xvi) the IGM spin temperature has not yet decoupled from the CMB; (xvii) the IGM spin temperature has not yet decoupled from the CMB; (xviii) the IGM spin temperature has not yet decoupled from the CMB; (xix) the IGM spin temperature has not yet decoupled from the CMB; (xx) the IGM spin temperature has not yet decoupled from the CMB; (xxi) the IGM spin temperature has not yet decoupled from the CMB; (xxii) the IGM spin temperature has not yet decoupled from the CMB; (xxiii) the IGM spin temperature has not yet decoupled from the CMB; (xxiv) the IGM spin temperature has not yet decoupled from the CMB; (xxv) the IGM spin temperature has not yet decoupled from the CMB; (xxvi) the IGM spin temperature has not yet decoupled from the CMB; (xxvii) the IGM spin temperature has not yet decoupled from the CMB; (xxviii) the IGM spin temperature has not yet decoupled from the CMB; (xxix) the IGM spin temperature has not yet decoupled from the CMB; (xxx) the IGM spin temperature has not yet decoupled from the CMB; (xxxi) the IGM spin temperature has not yet decoupled from the CMB; (xxxii) the IGM spin temperature has not yet decoupled from the CMB; (xxxiii) the IGM spin temperature has not yet decoupled from the CMB; (xxxiv) the IGM spin temperature has not yet decoupled from the CMB; (xxxv) the IGM spin temperature has not yet decoupled from the CMB; (xxxvi) the IGM spin temperature has not yet decoupled from the CMB; (xxxvii) the IGM spin temperature has not yet decoupled from the CMB; (xxxviii) the IGM spin temperature has not yet decoupled from the CMB; (xxxix) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the CMB; (xlv) the IGM spin temperature has not yet decoupled from the CMB; (xli) the IGM spin temperature has not yet decoupled from the CMB; (xlii) the IGM spin temperature has not yet decoupled from the CMB; (xliii) the IGM spin temperature has not yet decoupled from the CMB; (xiv) the IGM spin temperature has not yet decoupled from the
with figs 1 and 2 of Furlanetto & Loeb (2002) (although note that in Fig. 13, we plot \( \tau \) as a function of observed rather than intrinsic frequency). An entropy floor greatly reduces the cross-sectional area over which an observable absorption signal may be detected. If one detects significant mini-halo absorption over a large comoving patch along the line of sight to a radio source, one can place an upper limit on the entropy floor there. This in turn places an upper limit on the high-redshift X-ray background (from equation 8), and a lower limit on the redshift at which that patch was first reionized (since fossils from higher redshift have lower entropy, as in Fig. 2).

The chief driver of high-redshift 21-cm proposals has of course been to observe the IGM itself in 21-cm emission (Madau et al. 1997; Tozzi et al. 2000; Ciardi & Madau 2003). We note that if the filling factor of fossil H\( \text{II} \) regions is large, conditions are very favourable for such observations. Provided \( T_K \to T_K \) in the IGM, the only condition for the IGM to be seen in 21-cm emission against the CMB is for \( T_K \gg T_{\text{CMB}}(z_{\text{obs}}) \). This is certainly satisfied in fossil H\( \text{II} \) regions. It was previously thought that recoil heating from Ly\( \alpha \) photons could heat neutral regions (Madau et al. 1997), but a detailed calculation (Chen & Miralda-Escude 2003) shows the heating to be insufficient. In that case, a high-redshift X-ray background would be required to heat neutral regions. Such concerns are moot if a period of early reionization took place (as seems to be indicated by WMAP observations), followed by recombination: large tracts of warm, largely neutral gas would exist. Note, however, that brightness temperature fluctuations can only be detected if the contribution from unresolved radio point sources (Tozzi et al. 2000; Oh & Mack 2003) can be successfully removed by using spectral structure in frequency space.

5 EFFECTS ON GLOBAL REIONIZATION SCENARIOS

5.1 General considerations

Next we address the effect of the suppression of mini-halo formation on the global reionization history. The importance of the effect depends on: (i) the synchronization of the formation of mini-haloes \( t_{\text{sync}} \) relative to the typical lifetime \( t_{\text{ion}} \) of the ionizing source, which determines the fraction of mini-haloes subjected to feedback and (ii) on the recombination time \( t_{\text{rec}} \) relative to \( t_{\text{sync}} \), which determines how long the ionized regions last and thus the fraction of active (as opposed to fossil) ionized regions at any given time.

In the limit of \( t_{\text{sync}} \ll t_{\text{ion}} \), the fossil H\( \text{II} \) regions would appear only after all the mini-haloes had already formed, and hence the feedback would have no effect on the total amount of reionization by mini-haloes. In the absence of any other feedback effects, the IGM could then, in principle, be fully reionized by mini-haloes, given a high enough ionizing photon production efficiency (although would subsequently recombine).

In the opposite limit, \( t_{\text{sync}} \gg t_{\text{ion}} \), the contribution of mini-haloes to reionization will be strongly suppressed. We can obtain a rough estimate for the maximum fraction of the IGM that can be ionized by the mini-haloes. During the lifetime of its resident ionizing source, each mini-halo produces an ionized volume \( V_{\text{HII}} \), which will correspond to the total number \( N_{\gamma} \) of ionizing photons injected into the IGM, i.e. \( \bar{n}V \approx N_{\gamma} \), where \( \bar{n} \) is the mean hydrogen density.

To a good approximation, recombinations can be ignored in this phase, since massive stars have lifetimes shorter than the recombination time, \( t_{\text{rec}} \approx 3 \times 10^7 (C(1+z)/25)^{-1} \) yr (see Fig. 15). On the other hand, the recombination time is shorter than the Hubble time, \( t_{\text{rec}}/t_{\text{h urb}} \approx [C(1+z)/11]^{-3/2} \), so in general, at high redshift \( z > 10 \), and for clumping factors \( C > 1 \), the ionized volume will recombine once its driving source turns off.

As argued above, for the purpose of the mini-halo suppression, we may nevertheless simply imagine that the H\( \text{II} \) region never recombines. Ignoring recombinations inside the ionized region, this is equivalent to setting the size of the fossil H\( \text{II} \) region to be the maximum (comoving) size of the active H\( \text{II} \) region, reached at the time when the ionizing source turns off.

The evolution of the radii of the active and fossil H\( \text{II} \) regions are illustrated in Fig. 15 for a single \( 10^3 \) M\( _\odot \) metal-free star that turns on at \( z = 25 \). We follow the expansion of the ionization front \( R_f \) by solving the standard differential equation, taking into account ionizations, recombinations and the Hubble expansion (e.g. Shapiro & Giroux 1987; Cen & Haiman 2000). We assume a constant clumping factor of \( C = 1 \) (upper solid curve) or \( C = 10 \) (lower solid curve). We assume further that the metal-free star emits \( 40 \) 000 ionizing photons per stellar proton at a constant rate for \( \sim 3 \times 10^6 \) yr (Tumlinson & Shull 2000; Bromm et al. 2001b; Schaerer 2002). Once the ionizing source turns off, the solid curves show the formal solution for the evolution of the ionization front. The dashed curves show the size of the fossil H\( \text{II} \) region.

Next we can consider the evolution of the volume filling factor of the fossil H\( \text{II} \) regions from the ensemble of mini-haloes, and define the epoch \( z_f \) when the filling factor of these fossils reaches unity. As argued above, no new mini-haloes can form at \( z < z_f \). At this epoch, the global ionized fraction will be smaller than unity, because each fossil H\( \text{II} \) region has already partially recombined. Thus the global ionized fraction will be \( \sim \exp[-t_{\text{sync}}/t_{\text{ion}}] \), assuming that the filling factor of these fossils is large.

---

\( \text{C} = 1 \)

\( \text{C} = 10 \)

---

\[ R_f [\text{comoving Mpc}] \]

\[ z \]

---

\[ R_f (\text{comoving Mpc}) \]

\[ C = 1 \]

\[ C = 10 \]

---

\[ \text{Fig. 15.} \quad \text{The evolution of comoving radius of the ionized region around a single } 10^3 \text{ M}_\odot \text{ metal-free star that turns on at } z = 25, \text{ and is assumed to emit } 40,000 \text{ ionizing photons in } \sim 3 \times 10^6 \text{ yr. The upper and lower solid curves show solutions that include recombinations with } C = 1 \text{ and } 10. \text{ The ionized volume shrinks [or equivalently, the ionized fraction is decreased by the factor } (R_f/R_{\text{max}})^3] \text{ after the source turns off at } z \sim 24. \text{ The corresponding dashed curves show the radii of the fossil H II regions; these are assumed to stall at the maximum radius of the H II sphere and stay constant thereafter.} \]

---

\( \text{© 2003 RAS, MNRAS 346, 456–472} \)
sources typically turned off a time $t_{\text{sync}}$, and $t_{\text{rec}}$ is the average recombination time over the interval $t_{\text{sync}}$. For example, assuming that the fossil H II regions overlap at $z_f = 20$, with the typical sources born at $z = 25$, and the recombination time evaluated at $z = 22.5$ (with $C = 1$), we find $t_{\text{sync}} \approx t_{\text{rec}} \approx 5 \times 10^7$ yr, and a maximal ionized fraction of $\sim 30$ per cent.

### 5.2 Models for the reionization history

To refine the above considerations, next we compute the evolution of the global ionized fraction using a semi-analytical model adopted from Haiman & Holder (2003, hereafter HH03). For technical details, the reader is referred to that paper. Here we only briefly summarize the main ingredients of the model, and describe the modifications we made to include the suppression of the formation of new mini-haloes inside fossil H II regions. Feedback in actively ionized regions has already been considered in previous reionization models (Haiman & Loeb 1997, 1998; Haiman & Holder 2003; Wyithe & Loeb 2003). The main purpose of the new calculations presented here is to demonstrate that adding the additional feedback in the fossil H II regions has an important further impact on the reionization history.

In this model, we follow the volume filling fraction $F_{\text{HII}}$ of ionized regions, assuming that discrete ionized Strömgren spheres are being driven into the IGM by ionizing sources located in dark matter haloes. For simplicity, we ignore helium in this work. The dark matter halo mass function is adopted from Jenkins et al. (2001). We consider two distinct types of ionizing sources, located in haloes with virial temperatures of $10^5 < T_{\text{vir}} < 10^6$ K (mini-haloes) or $T_{\text{vir}} > 10^6$ K (large haloes). We assume that in mini-haloes, a fraction $f_s = 0.005$ of the baryons turn into metal-free stars (Bromm, Coppi & Larson 1999; Abel et al. 2000; Abel, Bryan & Norman 2002; Bromm et al. 2002), which produce $N_{\gamma} = 40000$ ionizing photons per baryon, and $f_{\text{esc}} = 100$ per cent of these photons escape into the IGM. The product $\epsilon = f_s f_{\text{esc}} N_{\gamma} = 200$ determines the size of the ionized regions around mini-haloes. For large haloes, we adopt $f_s = 0.1$, $f_{\text{esc}} = 20$ per cent and $N_{\gamma} = 4000$, for a total efficiency $\epsilon = 80$. In contrast with HH03, here we do not distinguish haloes with $10^5 < T_{\text{vir}} < 2 \times 10^5$ K from $T_{\text{vir}} > 2 \times 10^5$ K (see Dijkstra et al. 2003 for the reason why this distinction is probably unimportant).

As in HH03, we then follow the evolution of the global ionized fraction $F_{\text{HII}}(z)$ by summing the ionized volumes surrounding the individual dark matter haloes. We exclude the formation of mini-haloes in actively ionized regions (but assume larger haloes are impervious to photoionization feedback). This is accomplished by multiplying the rate of formation of new mini-haloes by the suppression factor $1 - F_{\text{HII}}$, as in equation (8) in Haiman & Holder (2003) (see also equation 11 in Wyithe & Loeb 2003 and section 3 in Haiman & Loeb 1997 for similar treatments of the actively ionized zones). The evolution of $F_{\text{HII}}$ in this model is shown by the upper thin solid curve in Fig. 16 (with $C = 1$ in the upper panel, and $C = 10$ in the lower panel). As the figure shows, in the absence of other feedback effects (see the discussion in Haiman et al. 2000; Haiman 2003 for other feedback effects), the mini-haloes could ionize the IGM in full by $z \sim 14$ ($C = 1$), or nearly fully by $z \sim 10$ (if clumping is assumed to be more significant, $C = 10$).

Next we compute $F_{\text{HII}}$ in the same model, except that we now also exclude the formation of new mini-haloes in fossil H II regions. This is easily accomplished in practice: the suppression factor $(1 - F_{\text{HII}})$ for the formation rate of mini-haloes is replaced by a factor $(1 - F_{\text{HII}}')$. Here the fossil filling factor $F_{\text{HII}}'$ is computed the same way as the ionized fraction $F_{\text{HII}}$, except that the individual ionized regions are assumed to follow the dashed curves from Fig. 15 rather than the solid curves. In Fig. 16, we show the evolution of the volume filling factor $F_{\text{HII}}'$ of the fossil H II regions (dashed curves), and the reionization history $F_{\text{HII}}(z)$ as the thick solid curves. Finally, for reference, we show $F_{\text{HII}}'(z)$ with mini-haloes completely excluded ($\epsilon = 0$) as the lower thin solid curves.

As apparent from Fig. 15, the exclusion of new mini-halo formation from the fossil H II regions causes a significant suppression of the total ionized fraction that can be reached by mini-haloes. Under the rather optimistic set of assumptions described by the thick solid curve in the upper panel, the maximum ionized fraction that can be reached is $\sim 40$ per cent. In reality, clumping is unlikely to be unity in the immediate vicinity ($\leq 100$ kpc) of the ionizing sources (Haiman, Abel & Madau 2001). The total electron scattering optical depth attributable to mini-haloes (the appropriately weighted integral between the thick solid curve and the lower thin solid curve) is $\tau = 0.07$ and 0.014 in the $C = 1$ and 10 cases, respectively. This makes it unlikely that mini-haloes can fully account for the large optical depth $\tau = 0.17$ measured by WMAP. Note the thick curve in the upper panel of Fig. 16 has a total $\tau = 0.2$ ($\tau = 0.07$ attributable to mini-haloes), and $\tau = 0.13$ to larger haloes, so that it is consistent with the WMAP measurement. Note that raising the efficiencies in mini-haloes would not increase the optical depth attributed to mini-haloes, since feedback then would set in earlier (we have explicitly verified that $\tau$ is approximately independent of efficiencies over a range of multiplicative factors 0.1–10 for $\epsilon$). Finally, note that the suppression considered here and shown in Fig. 16 provides a negative feedback in addition to the negative feedback expected from H$_2$ photodissociation (Haiman et al. 1997, 2000). In fact, as discussed above, pre-heating amplifies the effect of the H$_2$ photodissociative negative feedback.

Because an entropy floor essentially eliminates gas clumping due to mini-haloes (as shown in Section 4.1), the reionization
history is likely to more closely approximate the $C = 1$ case than the $C = 10$ case until low redshifts $z < 10$ (when gas clumping due to larger structures predominates). An entropy floor therefore has two counter-veiling effects on reionization: by suppressing star formation in mini-haloes, it reduces the comoving emissivity. However, by reducing gas clumping, it also reduces the photon budget required for reionization. It is interesting to note that the evolution of the filling factor is non-monotonic for the $C = 1$ case: feedback due to early reionization naturally produces a bump in the comoving emissivity, and a pause ensues before larger haloes (which can resist feedback) collapse. This is similar to the reionization histories derived by Cen (2003) and Wyithe & Loeb (2003), but is regulated by feedback from the entropy floor rather than a Population III to II transition due to a universal metallicity increase.

6 CONCLUSIONS

In this paper, motivated by the WMAP results, we have considered the feedback effect of early reionization/pre-heating on structure formation. This feedback effect is inevitable in any reionization scenario in which star formation throughout the Universe is not completely synchronized. Our principal conclusions are as follows.

(i) Fossil HII regions have a residual entropy floor after recombination and Compton cooling that is higher than the shock entropy for mini-haloes ($T_{\nu K} < 10^{4}$ K); thus, such haloes accrete gas isentropically. The IGM entropy depends primarily on the redshift and only weakly on the overdensity $\delta$; it is thus largely independent of the details of structure formation. For this reason, and also because it is conserved during adiabatic accretion or Hubble expansion, the gas entropy is a more fundamental variable to track than the temperature. We provide a simple analytic formula for the temperature (and hence entropy $K = T/n^{\alpha}$), in equations (33) and (35). An early X-ray background would also heat the entire IGM to similarly high adiabats. The entropy floor due to the latter would be much more spatially uniform.

(ii) We apply the entropy formalism used to calculate the effect of pre-heating on low-redshift galaxy clusters to the high-redshift mini-haloes. We obtained detailed gas density and pressure profiles, which we use to calculate the impact of pre-heating on the central density, accreted gas fraction, gas clumping factor and mini-halo baryonic mass function.

(iii) These quantities can then be used to calculate global effects of pre-heating. The collapsed gas fraction in mini-haloes falls by approximately two orders of magnitude, while the gas clumping factor falls to $C \rightarrow 1$, as for a uniform IGM. An entropy floor reduces the photon budget required for full reionization by approximately an order of magnitude, by reducing gas clumping and eliminating the need for mini-haloes to be photoevaporated before reionization can be completed.

(iv) However, an entropy floor does not necessarily promote early reionization: it also sharply reduces the comoving emissivity. By reducing the central gas densities in mini-haloes, pre-heating impedes H2 formation and cooling, and reduces the critical UV background required for H2 suppression by two to four orders of magnitude. Thus, once a comoving patch of the IGM is reionized, no subsequent star formation in mini-haloes can take place in that volume. The patch can only be reionized by more massive haloes $T > 10^{4}$ K, which can undergo atomic cooling. By furnishing an entropy floor, X-rays also suppress H2 formation. Thus, contrary to conventional wisdom, X-rays provide negative rather than positive feedback for early star formation. We note, however, the results of Ricotti et al. (2002b), who showed that gas in dense parts of filaments and in the interiors of galactic haloes can experience an enhancement of their H2 abundance, if they are engulfed by an HII front and the flux then turns off.

(v) Mini-haloes will not be observable in 21-cm emission/absorption in fossil HII regions. Thus, 21-cm observations provide an unusual probe of the topology of reionization: mini-haloes trace out regions of the IGM that have never been ionized. If mini-haloes are seen in large numbers, this places an upper limit on the filling factor of fossil HII regions and the X-ray background at that redshift.

(vi) We have computed the reionization histories as in HH03, but taking the feedback effect of an entropy floor (and reduction of gas clumping) into account. This is different from previous models of the reionization history that take into account the feedback only in actively ionized regions of the IGM. The strong additional feedback in fossil HII regions imply that HII fronts at high redshift never overlap, and global reionization at high redshift does not occur. This limits the contribution of mini-haloes to the reionization optical depth $\tau \sim 0.07$, almost independent of star formation efficiency in mini-haloes (if star formation is more efficient, feedback sets in earlier). Thus, the bulk of the optical depth observed by WMAP must come from more massive objects.

(vii) Strikingly, we obtain a double-peaked reionization history: an early peak in which the Universe is filled with fossil HII regions, followed by a pause before more massive haloes collapse which finally fully reionize the Universe. This is similar to ‘double reionization’ scenarios computed by other authors (e.g. Cen 2003; Wyithe & Loeb 2003), but one in which the comoving emissivity is regulated by gas entropy, rather than a Population III to II transition due to a universal metallicity increase. This last point deserves additional comment. A metallicity-regulated evolution of the emissivity requires that metal pollution is fairly spatially uniform. However, different parts of the IGM likely undergo the Population III to II transition at different epochs. Furthermore, an increase in metallicity does not necessarily result in a drop in the overall emissivity: metal line cooling probably results in greater star formation efficiency in haloes, since metals are not subject to internal UV photodissociation, unlike H2. The factor of $\sim 10$ drop in the H1 ionizing emissivity per stellar baryon could be overweighted by the increase in the total mass of stars formed. In comparison, we argue that an entropy-regulated transition is inevitable, and therefore more robust. This is an important conclusion, given the possibility that future CMB polarization studies will be able to distinguish among different reionization histories (HH03; Holder et al. 2003).

In this semi-analytic study, we essentially assumed a single value of the entropy floor at each redshift, for all haloes. In reality, as we discussed above, there should be large fluctuations in entropy, depending on the topology/history of reionization and the accretion/merger history of haloes. It would be interesting to study these effects in detail in three-dimensional numerical simulations. While it would be challenging to resolve the dynamics of gas in the cores of mini-haloes in large cosmological simulations, the problem appears well-suited to high-resolution simulations that zoom in to study the behaviour of a single collapsing halo (Bromm et al. 1999; Abel et al. 2000). It would be interesting to re-run existing simulations of the collapse of gas in mini-haloes, but adding an entropy floor in the initial conditions. Such studies could quantify more precisely the effects of an entropy floor in suppressing H2 formation and cooling in mini-haloes. In fact, since densities are lower and gas cooling is
reduced, it should be a computationally more tractable problem. With the aid of larger volumes, several global issues mentioned above could be addressed: for instance, a global, self-consistent study of feedback as in Machacek et al. (2001), but including the effects of both the entropy floor and UV feedback. Because of its strong effect on the evolution of the comoving emissivity, gas entropy acts as a self-regulating mechanism which probably has a strong influence in controlling the progress of reionization.

REFERENCES


Ciardi B., Madau P., in press (astro-ph/0303249)

Rosati P., Borgani S., Norman C., 2002, ARA&A


APPENDIX: ANALYTIC EXPRESSION FOR THE FINAL TEMPERATURE

It is useful to have an approximate analytic expression for the final temperature a parcel of gas cools down to, given the initial temperature $T_i$, redshift $z$, overdensity $\delta$ and length of time spent cooling $t$. This allows one to quickly estimate the effect of early reionization in different situations without evolving the full chemistry code. We develop such an expression in this Appendix.

At the redshifts and overdensities of interest, Compton cooling dominates by far. The most relevant physics is Compton cooling and hydrogen recombination:

$$\dot{T} = -\frac{8aT_i^7\sigma_T}{3mc} \left( x_e + \frac{x_r}{x_e} \right)$$

$$x_e = x_r^2n_a$$

Note that $(T - T_i) \approx T$ since $T \gg T_i$, and $\alpha \propto T^{-0.7}$ in the temperature range of interest. Assuming the gas is fully ionized $x_e = 1$ at some initial temperature $T_i$, we obtain the analytic solution:

$$T(x_e) = T_i \left( 1 + 1.4A \left[ \ln \left( \frac{1 + x_e}{x_e} \right) - \ln(2) \right] \right)^{-1/0.7}$$

where

$$A = \frac{4}{3} \frac{aT_i^7\sigma_T}{\alpha n(T_i) T_i} = \frac{\alpha n_i(T_i)}{\alpha n(T_i)} \left( \frac{T_i}{T} \right)^{0.7}$$

This gives the temperature as a function of the ionization fraction $x_e$. We therefore need to know the final ionization fraction. How can we estimate it? If the gas recombines isothermally at temperature $T'$, the ionization fraction is given by

$$x_e(t) = \frac{x_0}{1 + [t/t_{cool}(T')]}$$
where \( x_0 = 1 \) is the initial ionization fraction. If the gas cools as it recombines, substituting the instantaneous temperature \( T' \) into the expression would overestimate the speed of recombination and underestimate \( x_e \). We find that if we substitute \( T' = 2T \) and substitute this into equation (33), the solution of this non-linear equation for \( T_f \) is remarkably close to the full non-equilibrium solution (see points in Fig. 1). We have also verified that we obtain fairly accurate results for \( x_e(t) \). Equations (33) and (35) thus give \( T_f(z, \delta, T_i, t) \).

Our neglect of recombination line cooling fails in high-density regions. This leads to at most a factor of \(~2\) error in the final temperature, since recombination line cooling can cool gas down to at most \(~5000\) K. The entire discussion assumes that \( \text{H}_2 \) formation and cooling is not competitive with Compton cooling, which is generally true in low-density regions.

This paper has been typeset from a \TeX/\LaTeX file prepared by the author.