Sample Text for **The First Galaxies in the Universe**

Abraham Loeb
Department of Astronomy, Harvard University, 60 Garden St., Cambridge, MA 02138
Steven R. Furlanetto
Department of Physics & Astronomy, University of California Los Angeles, Los Angeles, CA 90095

Contents

1.2.1 Cosmological context: the expanding Universe

I. Introduction

When we look at our image reflected off a mirror at a distance of 1 meter, we see the way we looked 6.7 nanoseconds ago, the light travel time to the mirror and back. If the mirror is spaced 10^{19} cm $\simeq 3$ pc away, we will see the way we looked twenty one years ago. Light propagates at a finite speed, and so by observing distant regions, we are able to see what the Universe looked like in the past, a light travel time ago. The statistical homogeneity of the Universe on large scales guarantees that what we see far away is a fair statistical representation of the conditions that were present in our region of the Universe a long time ago.

This fortunate situation makes cosmology an empirical science. We do not need to guess how the Universe evolved. Using telescopes we can simply see how it appeared at earlier cosmic times. In principle, this allows the entire 13.7 billion year cosmic history of our universe to be reconstructed by surveying the galaxies and other sources of light to large distances. Since a greater distance means a fainter flux from a source of a fixed luminosity, the observation of the earliest sources of light requires the development of sensitive instruments and poses challenges to observers. As the universe expands, photon wavelengths get stretched as well. The factor by which the observed wavelength is increased (i.e. shifted towards the red) relative to the emitted one is denoted by $(1 + z)$, where z is the cosmological redshift. Astronomers use the known emission patterns of hydrogen and other chemical elements in the spectrum of each galaxy to measure z. This then implies that the universe has expanded by a factor of $(1 + z)$ in linear dimension since the galaxy emitted the observed light, and cosmologists can calculate the corresponding distance and cosmic age for the source galaxy. Large telescopes have allowed astronomers to observe faint galaxies that are so far away that we see them more than twelve billion years back in time. Thus, we know directly that galaxies were in existence as early as 850 million years after the Big Bang, at a redshift of $z \sim 6.5$ or higher.

We can in principle image the Universe only if it is transparent. Earlier than 400 000 years after the big bang, the cosmic hydrogen was broken into its constituent electrons and protons (i.e. "ionized") and the Universe was opaque to scattering by the free electrons in the dense plasma. Thus, telescopes cannot be used to electromagnetically image the infant Universe at earlier times (or redshifts $> 10^3$). The earliest possible image of the Universe was recorded by the COBE and WMAP satellites, which measured the temperature distribution of the cosmic microwave background (CMB) on the sky.

The CMB, the relic radiation from the hot, dense beginning of the universe, is indeed another major probe of observational cosmology. The universe cools as it expands, so it was initially far denser and hotter than it is today. For hundreds of thousands of years the cosmic gas consisted of a plasma of free protons and electrons, and a slight mix of light nuclei, sustained by the intense thermal motion of these particles. Just like the plasma in our own Sun, the ancient cosmic plasma emitted and scattered a strong field of visible and ultraviolet photons. About 400 000 years after the Big Bang the temperature of the universe dipped for the first time below a few thousand degrees Kelvin. The protons and electrons were now moving slowly enough that they could attract each other and form hydrogen atoms, in a process known as cosmic recombination. With the scattering of the energetic photons now much reduced, the photons continued traveling in straight lines, mostly undisturbed except that cosmic expansion has redshifted their wavelength into the microwave

regime today. The emission temperature of the observed spectrum of these CMB photons is the same in all directions to one part in 100 000, which reveals that conditions were nearly uniform in the early universe.

It was just before the moment of cosmic recombination (when matter started to dominate in energy density over radiation) that gravity started to amplify the tiny fluctuations in temperature and density observed in the CMB data. Regions that started out slightly denser than average began to contract because the gravitational forces were also slightly stronger than average in these regions. Eventually, after hundreds of millions of years of contraction, the overdense regions stopped expanding, turned around, and eventually collapsed to make bound objects such as galaxies. The gas within these collapsed objects cooled and fragmented into stars. This process, however, would have taken too long to explain the abundance of galaxies today, if it involved only the observed cosmic gas. Instead, gravity is strongly enhanced by the presence of dark matter – an unknown substance that makes up the vast majority (83%) of the cosmic density of matter. The motion of stars and gas around the centers of nearby galaxies indicates that each is surrounded by an extended mass of dark matter, and so dynamically-relaxed dark matter concentrations are generally referred to as "halos".

According to the standard cosmological model, the dark matter is cold (abbreviated as CDM), i.e., it behaves as a collection of collisionless particles that started out at matter domination with negligible thermal velocities and have evolved exclusively under gravitational forces. The model explains how both individual galaxies and the large-scale patterns in their distribution originated from the small initial density fluctuations. On the largest scales, observations of the present galaxy distribution have indeed found the same statistical patterns as seen in the CMB, enhanced as expected by billions of years of gravitational evolution. On smaller scales, the model describes how regions that were denser than average collapsed due to their enhanced gravity and eventually formed gravitationally-bound halos, first on small spatial scales and later on larger ones. In this hierarchical model of galaxy formation, the small galaxies formed first and then merged or accreted gas to form larger galaxies. At each snapshot of this cosmic evolution, the abundance of collapsed halos, whose masses are dominated by dark matter, can be computed from the initial conditions using numerical simulations. The common understanding of galaxy formation is based on the notion that stars formed out of the gas that cooled and subsequently condensed to high densities in the cores of some of these halos.

Gravity thus explains how some gas is pulled into the deep potential wells within dark matter halos and forms the galaxies. One might naively expect that the gas outside halos would remain mostly undisturbed. However, observations show that it has not remained neutral (i.e., in atomic form) but was largely ionized by the UV radiation emitted by the galaxies. The diffuse gas pervading the space outside and between galaxies is referred to as the intergalactic medium (IGM). For the first hundreds of millions of years after cosmological recombination, the so-called cosmic "dark ages", the universe was filled with diffuse atomic hydrogen. As soon as galaxies formed, they started to ionize diffuse hydrogen in their vicinity. Within less than a billion years, most of the IGM was re-ionized. We have not yet imaged the cosmic dark ages before the first galaxies had formed. One of the frontiers in current cosmological studies aims to study the cosmic epoch of reionization and the first generation of galaxies that triggered it.

II. Preliminaries

The modern physical description of the Universe as a whole can be traced back to Einstein, who assumed for simplicity the so-called "cosmological principle": that the distribution of matter and energy is homogeneous and isotropic on the largest scales. Today isotropy is well established for the distribution of faint radio sources, optically-selected galaxies, the X-ray background, and most importantly the CMB. The constraints on homogeneity are less strict, but a cosmological model in which the Universe is isotropic but significantly inhomogeneous in spherical shells around our special location, is also excluded.

In General Relativity, the metric for a space which is spatially homogeneous and isotropic is the Friedman-Robertson-Walker metric, which can be written in the form

$$ds^2 = c^2 dt^2 - a^2(t) \left[\frac{dR^2}{1 - k\,R^2} + R^2 \left(d\theta^2 + \sin^2\theta\, d\phi^2 \right) \right] , \tag{1}$$

where c is the speed of light, $a(t)$ is the cosmic scale factor which describes expansion in time t, and (R, θ, ϕ) are spherical comoving coordinates. The constant k determines the geometry of the metric; it is positive in a closed Universe, zero in a flat Universe, and negative in an open Universe. Observers at rest remain at rest, at fixed (R, θ, ϕ), with their physical separation increasing with time in proportion to $a(t)$. A given observer sees a nearby observer at physical distance D receding at the Hubble velocity $H(t)D$, where the Hubble constant at time t is $H(t) = d\,a(t)/dt$. Light emitted by a source at time t is observed at $t = 0$ with a redshift $z = 1/a(t) - 1$, where we set $a(t = 0) \equiv 1$ for convenience.

The Einstein field equations of General Relativity yield the Friedmann equation

$$H^2(t) = \frac{8\pi G}{3}\rho - \frac{k}{a^2} \ , \tag{2}$$

which relates the expansion of the Universe to its matter-energy content. The constant k determines the geometry of the universe; it is positive in a closed universe, zero in a flat universe, and negative in an open universe. For each component of the energy density ρ, with an equation of state $p = p(\rho)$, the density ρ varies with $a(t)$ according to the thermodynamic relation

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

With the critical density

$$\rho_C(t) \equiv \frac{3H^2(t)}{8\pi G} \tag{4}$$

defined as the density needed for $k = 0$, we define the ratio of the total density to the critical density as

$$\Omega \equiv \frac{\rho}{\rho_C} \ . \tag{5}$$

With Ω_m, Ω_Λ, and Ω_r denoting the present contributions to Ω from matter (including cold dark matter as well as a contribution Ω_b from ordinary matter ["baryons"] made of protons and neutrons), vacuum density (cosmological constant), and radiation, respectively, the Friedmann equation becomes

$$\frac{H(t)}{H_0} = \left[\frac{\Omega_m}{a^3} + \Omega_\Lambda + \frac{\Omega_r}{a^4} + \frac{\Omega_k}{a^2}\right] \ , \tag{6}$$

where we define H_0 and $\Omega_0 = \Omega_m + \Omega_\Lambda + \Omega_r$ to be the present values of H and Ω, respectively, and we let

$$\Omega_k \equiv -\frac{k}{H_0^2} = 1 - \Omega_m. \tag{7}$$

In the particularly simple Einstein-de Sitter model ($\Omega_m = 1$, $\Omega_\Lambda = \Omega_r = \Omega_k = 0$), the scale factor varies as $a(t) \propto t^{2/3}$. Even models with non-zero Ω_Λ or Ω_k approach the Einstein-de Sitter scaling-law at high redshift, i.e. when $(1+z) \gg |{\Omega_m}^{-1} - 1|$ (as long as Ω_r can be neglected). In this moderately high-z regime the age of the Universe is

$$t \approx \frac{2}{3H_0\sqrt{\Omega_m}}\,(1+z)^{-3/2} \ \approx 10^9 \mathrm{yr}\left(\frac{1+z}{7}\right)^{-3/2} . \tag{8}$$

Recent observations confine the standard set of cosmological parameters to a relatively narrow range. In particular, we seem to live in a universe dominated by a cosmological constant (Λ) and cold dark matter, or in short a ΛCDM cosmology (with Ω_k so small that it is usually assumed to equal zero) with an approximately scale-invariant primordial power spectrum of density fluctuations, i.e., $n \approx 1$ where the initial power spectrum is $P(k) = |\delta_{\mathbf{k}}|^2 \propto k^n$ in terms of the wavenumber k of the Fourier modes $\delta_{\mathbf{k}}$ (see § below). Also, the Hubble constant today is written as $H_0 = 100h$ km s^{-1}Mpc^{-1} in terms of h, and the overall normalization of the power spectrum is specified in terms of σ_8, the root-mean-square amplitude of mass fluctuations in spheres of radius 8 h^{-1} Mpc. For example, the best-fit cosmological parameters matching the WMAP data together with large-scale gravitational lensing observations are $\sigma_8 = 0.826$, $n = 0.953$, $h = 0.687$, $\Omega_m = 0.299$, $\Omega_\Lambda = 0.701$ and $\Omega_b = 0.0478$.

2.2 Growth of linear perturbations

As noted in the Introduction, observations of the CMB show that the universe at cosmic recombination (redshift $z \sim 10^3$) was remarkably uniform apart from spatial fluctuations in the energy density and in the gravitational potential of roughly one part in $\sim 10^5$. The primordial inhomogeneities in the density distribution grew over time and eventually led to the formation of galaxies as well as galaxy clusters and large-scale structure. In the early stages of this growth, as long as the density fluctuations on the relevant scales were much smaller than unity, their evolution can be understood with a linear perturbation analysis.

As before, we distinguish between fixed and comoving coordinates. Using vector notation, the fixed coordinate $\mathbf{r}$ corresponds to a comoving position $\mathbf{x} = \mathbf{r}/a$. In a homogeneous Universe with density ρ, we describe the cosmological expansion in terms of an ideal pressureless fluid of particles each of which is at

fixed $\mathbf{x}$, expanding with the Hubble flow $\mathbf{v} = H(t)\mathbf{r}$ where $\mathbf{v} = d\mathbf{r}/dt$. Onto this uniform expansion we impose small perturbations, given by a relative density perturbation

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{r})}{\bar{\rho}} - 1 \ , \tag{9}$$

where the mean fluid density is $\bar{\rho}$, with a corresponding peculiar velocity $\mathbf{u} \equiv \mathbf{v} - H\mathbf{r}$. Then the fluid is described by the continuity and Euler equations in comoving coordinates:

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot [(1+\delta)\mathbf{u}] = 0 \tag{10}$$

$$\frac{\partial \mathbf{u}}{\partial t} + H\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{a}\nabla\phi \ . \tag{11}$$

The potential ϕ is given by the Poisson equation, in terms of the density perturbation:

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \ . \tag{12}$$

This fluid description is valid for describing the evolution of collisionless cold dark matter particles until different particle streams cross. This "shell-crossing" typically occurs only after perturbations have grown to become non-linear, and at that point the individual particle trajectories must in general be followed. Similarly, baryons can be described as a pressureless fluid as long as their temperature is negligibly small, but non-linear collapse leads to the formation of shocks in the gas.

For small perturbations $\delta \ll 1$, the fluid equations can be linearized and combined to yield

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta \ . \tag{13}$$

This linear equation has in general two independent solutions, only one of which grows with time. Starting with random initial conditions, this "growing mode" comes to dominate the density evolution. Thus, until it becomes non-linear, the density perturbation maintains its shape in comoving coordinates and grows in proportion to a growth factor $D(t)$. The growth factor in the matter-dominated era is given by

$$D(t) \propto \frac{\left(\Omega_\Lambda a^3 + \Omega_k a + \Omega_m\right)^{1/2}}{a^{3/2}} \int_0^a \frac{a'^{3/2}\, da'}{\left(\Omega_\Lambda a'^3 + \Omega_k a' + \Omega_m\right)^{3/2}} \ , \tag{14}$$

where we neglect Ω_r when considering halos forming in the matter-dominated regime at $z \ll 10^4$. In the Einstein-de Sitter model (or, at high redshift, in other models as well) the growth factor is simply proportional to $a(t)$.

The spatial form of the initial density fluctuations can be described in Fourier space, in terms of Fourier components

$$\delta_{\mathbf{k}} = \int d^3x\, \delta(x) e^{-i\mathbf{k}\cdot\mathbf{x}} \ . \tag{15}$$

Here we use the comoving wave-vector $\mathbf{k}$, whose magnitude k is the comoving wavenumber which is equal to 2π divided by the wavelength. The Fourier description is particularly simple for fluctuations generated by inflation. Inflation generates perturbations given by a Gaussian random field, in which different $\mathbf{k}$-modes are statistically independent, each with a random phase. The statistical properties of the fluctuations are determined by the variance of the different $\mathbf{k}$-modes, and the variance is described in terms of the power spectrum $P(k)$ as follows:

$$\langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \rangle = (2\pi)^3 P(k)\, \delta^{(3)}\left(\mathbf{k} - \mathbf{k}'\right) \ , \tag{16}$$

where $\delta^{(3)}$ is the three-dimensional Dirac delta function. The gravitational potential fluctuations are sourced by the density fluctuations through Poisson's equation.

In standard models, inflation produces a primordial power-law spectrum $P(k) \propto k^n$ with $n \sim 1$. Perturbation growth in the radiation-dominated and then matter-dominated Universe results in a modified final power spectrum, characterized by a turnover at a scale of order the horizon cH^{-1} at matter-radiation equality, and a small-scale asymptotic shape of $P(k) \propto k^{n-4}$. The overall amplitude of the power spectrum is not specified by current models of inflation, and it is usually set by comparing to the observed CMB temperature fluctuations or to local measures of large-scale structure.

Since density fluctuations may exist on all scales, in order to determine the formation of objects of a given size or mass it is useful to consider the statistical distribution of the smoothed density field. Using a window function $W(\mathbf{r})$ normalized so that $\int d^3r\, W(\mathbf{r}) = 1$, the smoothed density perturbation field, $\int d^3r \delta(\mathbf{x}) W(\mathbf{r})$, itself follows a Gaussian distribution with zero mean. For the particular choice of a spherical top-hat, in

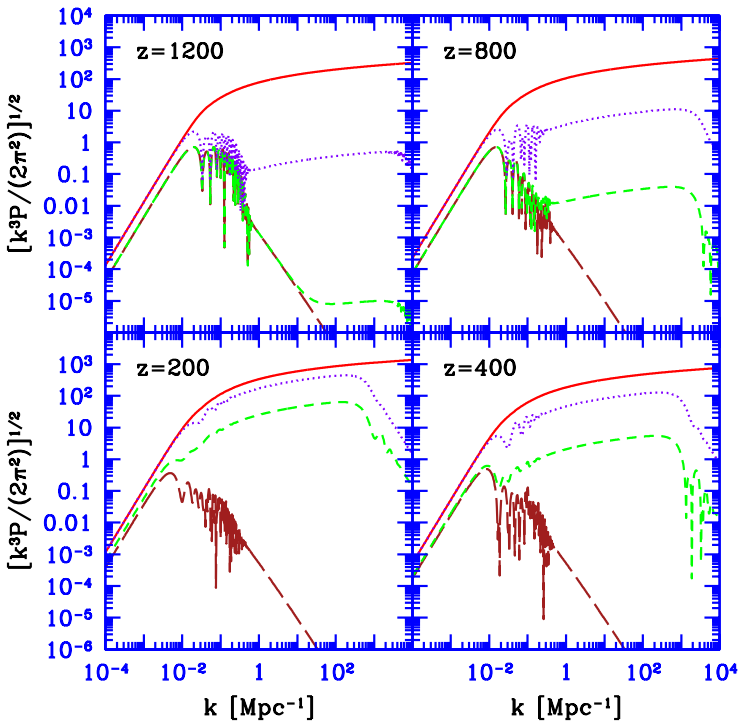

Figure 1: Power spectra of density and temperature fluctuations vs. comoving wavenumber, at redshifts 1200, 800, 400, and 200. We consider fluctuations in the CDM density (solid curves), baryon density (dotted curves), baryon temperature (short-dashed curves), and photon temperature (long-dashed curves).

which W =constant in a sphere of radius R and is zero outside, the smoothed perturbation field measures the fluctuations in the mass in spheres of radius R. The normalization of the present power spectrum is often specified by the value of $\sigma_8 \equiv \sigma(R = 8h^{-1}\mathrm{Mpc})$. For the top-hat, the smoothed perturbation field is denoted δ_R or δ_M, where the mass M is related to the comoving radius R by $M = 4\pi\rho_m R^3/3$, in terms of the current mean density of matter ρ_m. The variance $\langle\delta_M\rangle^2$ is

$$\sigma^2(M) = \sigma^2(R) = \int_0^\infty \frac{dk}{2\pi^2}\, k^2 P(k) \left[\frac{3j_1(kR)}{kR}\right]^2 , \tag{17}$$

where $j_1(x) = (\sin x - x\cos x)/x^2$. The function $\sigma(M)$ plays a crucial role in estimates of the abundance of collapsed objects, as we describe later.

Different physical processes contributed to the perturbation growth. In the absence of other influences, gravitational forces due to density perturbations imprinted by inflation would have driven parallel perturbation growth in the dark matter, baryons and photons. However, since the photon sound speed is of order the speed of light, the radiation pressure produced sound waves on a scale of order the cosmic horizon and suppressed sub-horizon perturbations in the photon density. The baryonic pressure similarly suppressed perturbations in the gas below the (much smaller) so-called baryonic *Jeans* scale. Since the formation of hydrogen at recombination had decoupled the cosmic gas from its mechanical drag on the CMB, the baryons subsequently began to fall into the pre-existing gravitational potential wells of the dark matter.

Spatial fluctuations developed in the gas temperature as well as in the gas density. Both the baryons and the dark matter were affected on small scales by the temperature fluctuations through the gas pressure. Compton heating due to scattering of the residual free electrons (constituting a fraction $\sim 10^{-4}$) with the CMB photons remained effective, keeping the gas temperature fluctuations tied to the photon temperature fluctuations, even for a time after recombination. The growth of linear perturbations can be calculated with the standard CMBFAST code (http://www.cmbfast.org), after a modification to account for the fact that the speed of sound of the gas also fluctuates spatially.

The magnitude of the fluctuations in the CDM and baryon densities, and in the baryon and photon temperatures, is shown in Figure 1, in terms of the dimensionless combination $[k^3P(k)/(2\pi^2)]^{1/2}$, where $P(k)$ is the corresponding power spectrum of fluctuations in terms of the comoving wavenumber k of each Fourier mode. After recombination, two main drivers affect the baryon density and temperature fluctuations, namely, the thermalization with the CMB and the gravitational force that attracts the baryons to the dark

matter potential wells. As shown in the figure, the density perturbations in all species grow together on scales where gravity is unopposed, outside the horizon (i.e., at $k < 0.01$ Mpc^{-1} at $z \sim 1000$). At $z = 1200$ the perturbations in the baryon-photon fluid oscillate as acoustic waves on scales of order the sound horizon ($k \sim 0.01$ Mpc^{-1}), while smaller-scale perturbations in both the photons and baryons are damped by photon diffusion and the drag of the diffusing photons on the baryons. On sufficiently small scales the power spectra of baryon density and temperature roughly assume the shape of the dark matter fluctuations (except for the gas-pressure cutoff at the very smallest scales), due to the effect of gravitational attraction on the baryon density and of the resulting adiabatic expansion on the gas temperature. After the mechanical coupling of the baryons to the photons ends at $z \sim 1000$, the baryon density perturbations gradually grow towards the dark matter perturbations because of gravity. Similarly, after the thermal coupling ends at $z \sim 200$, the baryon temperature fluctuations are driven by adiabatic expansion towards a value of 2/3 of the density fluctuations. As the figure shows, by $z = 200$ the baryon infall into the dark matter potentials is well advanced and adiabatic expansion is becoming increasingly important in setting the baryon temperature.

6.2 21-cm absorption or emission

6.2.1 Atomic physics

The fundamental quantity of radiative transfer is the *brightness* (or *specific intensity*) I_ν of a ray emerging from a cloud at frequency ν, or its angle-averaged form $J_\nu = \int I_\nu \mathrm{d}\Omega/4\pi$. This conventionally expresses the energy carried by rays traveling along a given direction, per unit area, frequency, solid angle, and time; it thus normally has dimensions ergs s^{-1} cm^{-2} sr^{-1} Hz^{-1} However, for many applications of radiative transfer in an expanding universe, the units cm^{-2} s^{-1} Hz^{-1} sr^{-1} are more convenient, because photon number is conserved during the expansion but energy is not.

For convenience, we will quantify I_ν by the equivalent *brightness temperature*, $T_b(\nu)$, required of a blackbody radiator (with spectrum B_ν) such that $I_\nu = B_\nu(T_b)$. Throughout the range of frequencies and temperatures relevant to the 21 cm line, the Rayleigh-Jeans formula is an excellent approximation to the Planck curve, so that $T_b(\nu) \approx I_\nu\, c^2/2k_B\nu^2$, where is c is the speed of light and k_B is Boltzmann's constant.

We will be almost exclusively interested in the brightness temperature of the HI 21 cm line, which has rest frequency $\nu_0 = 1420.4057$ MHz. Because of the cosmological redshift, the emergent brightness $T_b'(\nu_0)$ measured in a cloud's comoving frame at redshift z creates an apparent brightness at the Earth of $T_b(\nu) = T_b'(\nu_0)/(1+z)$, where the observed frequency is $\nu = \nu_0/(1+z)$. Similarly, the brightness temperature of the CMB in a comoving frame at redshift z scales from the presently observed value of $T_\gamma(0) = 2.73$ K to $T_\gamma'(z) = 2.73\ (1+z)$ K.

The radiative transfer equation for a spectral line reads,

$$\frac{dI_\nu}{ds} = \frac{\phi(\nu) h\nu}{4\pi} \left[n_1 A_{10} - (n_0 B_{01} - n_1 B_{10})\, I_\nu \right], \tag{18}$$

where ds is a line element, $\phi(\nu)$ is the line profile function normalized by $\int \phi(\nu) d\nu = 1$ (with an amplitude of order the inverse of the frequency width of the line), subscripts 0 and 1 denote the lower and upper levels, $n_{0,1}$ denotes the number density of atoms at the different levels, and A and B are the Einstein coefficients for the transition between these levels. We can then make use of the standard relations: $B_{10} = (g_0/g_1)B_{01}$ and $B_{01} = (g_1/g_0)A_{10}\Pi/I_\nu$, where g is the spin degeneracy factor of each state and Π is the photon occupation number. For the 21cm transition, $A_{10} = 2.85 \times 10^{-15}\mathrm{s}^{-1}$ and $g_1/g_0 = 3$.

In the Rayleigh-Jeans limit, the equation of radiative transfer along a line of sight through a cloud of uniform excitation temperature T_S implies that the emergent brightness at frequency ν is

$$T_b'(\nu) = T_{\rm ex}(1 - e^{-\tau_\nu}) + T_R'(\nu) e^{-\tau_\nu} \tag{19}$$

where the *optical depth* $\tau_\nu \equiv \int \mathrm{d}s\, \alpha_\nu$ is the integral of the absorption coefficient α_ν (which is obtained from the right-hand-side of Eq. 18) along the ray through the cloud, and T_R' is the brightness of the background radiation field incident on the cloud along the ray.

The relative populations of hydrogen atoms in the two spin states define the so-called spin temperature, T_S, through the relation

$$\frac{n_1}{n_0} = 3\, e^{-T_\star/T_S} \tag{20}$$

where $T_\star \equiv E_{10}/k_B = 0.068$ K is the equivalent temperature of the transition energy. Because all astrophysical applications have $T_S \gg T_*$, approximately three of four atoms find themselves in the excited state. As a result, the absorption coefficient must include a correction for stimulated emission (and hence it depends

on T_S as well). Moreover, in the regime of interest, $T_\star$ is also much smaller than the spin temperature T_S, and so all related exponentials can be expanded to leading order.

The optical depth of a cloud of hydrogen is then:

$$\tau_\nu = \int \mathrm{d}s\, \sigma_{01}\, (1 - e^{-E_{10}/k_B T_S})\, \phi(\nu)\, n_0 \tag{21}$$

$$\approx \sigma_{01} \left(\frac{h\nu}{k_B T_S}\right) \left(\frac{N_{\rm HI}}{4}\right) \phi(\nu), \tag{22}$$

where

$$\sigma_{01} \equiv \frac{3c^2 A_{10}}{8\pi\nu^2}, \tag{23}$$

$N_{\rm HI}$ is the column density of HI (here the factor 1/4 accounts for the fraction of HI atoms in the hyperfine singlet state). The second factor in equation (21) accounts for stimulated emission.

In general, the line shape $\phi(\nu)$ includes natural, thermal, and pressure broadening, as well as bulk motion (which increases the effective Doppler spread). The most important application is to IGM gas expanding uniformly with the Hubble flow. Then the velocity broadening of a region of linear dimension s will be $\Delta V \sim sH(z)$ so that $\phi(\nu) \sim c/[sH(z)\nu]$. The column density along such a segment depends on the neutral fraction $x_{\rm HI}$ of hydrogen, so $N_{\rm HI} = x_{\rm HI} n_H(z)\ s$. A more exact calculation yields, with equation (21), an expression for the 21 cm optical depth of the diffuse IGM,

$$\tau_{\nu_0} = \frac{3}{32\pi} \frac{hc^3 A_{10}}{k_B T_S \nu_0^2} \frac{x_{\rm HI} n_H}{(1+z)\,(\mathrm{d}v_\parallel/\mathrm{d}r_\parallel)} \tag{24}$$

$$\approx 0.0092\,(1+\delta)\,(1+z)^{3/2}\, \frac{x_{\rm HI}}{T_S} \left[\frac{H(z)/(1+z)}{\mathrm{d}v_\parallel/\mathrm{d}r_\parallel}\right], \tag{25}$$

where in the second equality T_S is in degrees Kelvin. Here the factor $(1+\delta)$ is the fractional overdensity of baryons and $\mathrm{d}v_\parallel/\mathrm{d}r_\parallel$ is the gradient of the proper velocity along the line of sight, including both the Hubble expansion and the peculiar velocity. In the second line, we have substituted the velocity $H(z)/(1+z)$ appropriate for the uniform Hubble expansion at high redshifts.

The most important application of equation (19) is observing high-redshift hydrogen clouds against the CMB. Thus we hope to measure

$$\delta T_b(\nu) = \frac{T_S - T_\gamma(z)}{1+z}(1 - e^{-\tau_{\nu_0}}) \approx \frac{T_S - T_\gamma(z)}{1+z}\, \tau_{\nu_0} \tag{26}$$

$$\approx 9\, x_{\rm HI}(1+\delta)\,(1+z)^{1/2} \left[1 - \frac{T_\gamma(z)}{T_S}\right] \left[\frac{H(z)/(1+z)}{\mathrm{d}v_\parallel/\mathrm{d}r_\parallel}\right]\ \mathrm{mK}. \tag{27}$$

Note that δT_b saturates if $T_S \gg T_\gamma$, but it can become arbitrarily large (and negative) if $T_S \ll T_\gamma$. The observability of the 21 cm transition therefore hinges on the spin temperature.

Three competing processes determine T_S: (1) absorption of CMB photons (as well as stimulated emission); (2) collisions with other hydrogen atoms, free electrons, and protons; and (3) scattering of UV photons. We let C_{10} and P_{10} be the de-excitation rates (per atom) from collisions and UV scattering, respectively, with C_{01} and P_{01} be the corresponding excitation rates. The equilibrium spin temperature is then determined by

$$n_1\,(C_{10} + P_{10} + A_{10} + B_{10} I_{\rm CMB}) = n_0\,(C_{01} + P_{01} + B_{01} I_{\rm CMB})\,, \tag{28}$$

where B_{01} and B_{10} are the appropriate Einstein coefficients and $I_{\rm CMB}$ is the energy flux of CMB photons. With the Rayleigh-Jeans approximation, equation (28) can be rewritten as

$$T_S^{-1} = \frac{T_\gamma^{-1} + x_c T_K^{-1} + x_\alpha T_c^{-1}}{1 + x_c + x_\alpha}, \tag{29}$$

where x_c and x_α are coupling coefficients for collisions and UV scattering, respectively, and T_K is the gas kinetic temperature. Here we have used detailed balance through the relation

$$\frac{C_{01}}{C_{10}} = \frac{g_1}{g_0} e^{-T_\star/T_K} \approx 3\left(1 - \frac{T_\star}{T_K}\right). \tag{30}$$

We have then *defined* the effective color temperature of the UV radiation field T_c via

$$\frac{P_{01}}{P_{10}} \equiv 3\left(1 - \frac{T_\star}{T_c}\right). \tag{31}$$

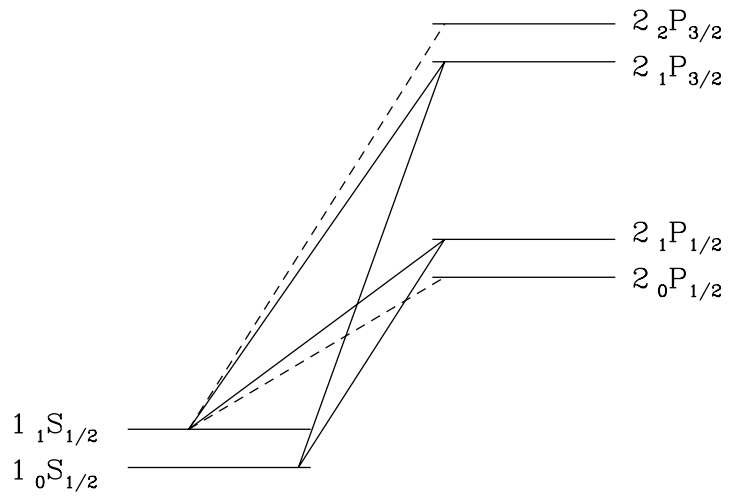

Figure 2: Level diagram illustrating the Wouthuysen-Field effect. We show the hyperfine splittings of the $1S$ and $2P$ levels of neutral hydrogen. The solid lines label transitions that mix the ground state hyperfine levels, while the dashed lines label complementary transitions that do not participate in mixing.

Collisional excitation and de-excitation of the hyperfine levels dominate in dense gas. The coupling coefficient for species i is

$$x_c^i \equiv \frac{C_{10}^i}{A_{10}} \frac{T_\star}{T_\gamma} = \frac{n_i \, \kappa_{10}^i}{A_{10}} \frac{T_\star}{T_\gamma}, \tag{32}$$

where κ_{10}^i is the rate coefficient for spin de-excitation in collisions with that species (with units of $\mathrm{cm}^3\ \mathrm{s}^{-1}$). The total x_c is the sum over all i, including hydrogen-hydrogen, hydrogen-electron, and hydrogen-proton collisions. All of these rate coefficients can be computed quite precisely (to $< 10\%$ error) using the relevant quantum mechanical scattering cross sections; hydrogen-hydrogen collisions generally dominate unless the ionized fraction is larger than a few percent.

The second coupling mechanism is scattering of Lyman-series photons through the IGM, the so-called "Wouthuysen-Field effect."[1] It is illustrated in Figure 2, where we have drawn the hyperfine sublevels of the $1S$ and $2P$ states of HI. Suppose a hydrogen atom in the hyperfine singlet state absorbs a Lyα photon. The electric dipole selection rules allow $\Delta F = 0, 1$ except that $F = 0 \to 0$ is prohibited (here F is the total angular momentum of the atom). Thus the atom will jump to either of the central $2P$ states. However, these rules allow this state to decay to the ${}_1S_{1/2}$ triplet level.[2] Thus atoms can change hyperfine states through the absorption and spontaneous re-emission of a Lyα photon. This is analogous to the well-known "Raman scattering" process, which often determines the level populations of metastable atomic states, except that in this case the atom undergoes a real (rather than virtual) transition to the $2P$ state.

We will provide a relatively simple and intuitive treatment of this process; more detailed calculations require consideration of the line structure. The Wouthuysen-Field coupling must depend on the total rate (per atom) at which Lyα photons are scattered within the gas,

$$P_\alpha = 4\pi\chi_\alpha \int \mathrm{d}\nu \, J_\nu(\nu)\phi_\alpha(\nu), \tag{33}$$

where $\sigma_\nu \equiv \chi_\alpha \phi_\alpha(\nu)$ is the local absorption cross section, $\chi_\alpha \equiv (\pi\, e^2/m_e\, c) f_\alpha$, $f_\alpha = 0.4162$ is the oscillator strength of the Lyα transition, $\phi_\alpha(\nu)$ is the Lyα absorption profile, and J_ν is the angle-averaged specific

[1] As a guide to the English-speaking reader, "Wouthuysen" is pronounced as roughly "Vowt-how-sen," although in reality the "uy" construction is a diphthong with no precise counterpart in English.

[2] Here we use the notation ${}_F L_J$, where L and J are the orbital and total angular momentum of the electron.

intensity of the background radiation field (by number, not energy). In the simplest approximation, we take J_ν to be constant across the line.

Our goal is to relate this total scattering rate P_α to the indirect de-excitation rate P_{10}. We first label the $1S$ and $2P$ hyperfine levels a–f, in order of increasing energy, and let A_{ij} and B_{ij} be the spontaneous emission and absorption coefficients for transitions between these levels. We write the background flux at the frequency corresponding to the $i \to j$ transition as J_{ij}. Then

$$P_{01} \propto B_{\rm ad} J_{\rm ad} \frac{A_{\rm db}}{A_{\rm da} + A_{\rm db}} + B_{\rm ae} J_{\rm ae} \frac{A_{\rm eb}}{A_{\rm ea} + A_{\rm eb}}. \tag{34}$$

The first term contains the probability for an a→d transition ($B_{\rm ad} J_{\rm ad}$), together with the probability for the subsequent decay to terminate in state b; the second term is the same for transitions to and from state e. Next we need to relate the individual A_{ij} to $A_\alpha = 6.25 \times 10^8$ Hz, the total Lyα spontaneous emission rate (averaged over all the hyperfine sublevels). This can be accomplished using a sum rule stating that the sum of decay intensities ($g_i A_{ij}$) for transitions from a given nFJ to all the $n'J'$ levels (summed over F') is proportional to $2F + 1$; the relative strengths of the permitted transitions are then (1, 1, 2, 2, 1, 5), where we have ordered the lines (bc, ad, bd, ae, be, bf) and the two letters represent the initial and final states. With our assumption that the background radiation field is constant across the individual hyperfine lines, we then find $P_{10} = (4/27)P_\alpha$.

The coupling coefficient x_α may then be written

$$x_\alpha = \frac{4P_\alpha}{27A_{10}} \frac{T_\star}{T_\gamma} = S_\alpha \frac{J_\nu}{J_\nu^c}, \tag{35}$$

where in the second equality we have again taken J_ν to be constant around the line and set the fiducial constant $J_\nu^c \equiv 1.165 \times 10^{-10}[(1+z)/20]$ cm^{-2} s^{-1} Hz^{-1} sr^{-1}. A more detailed calculation shows that the spectrum is typically suppressed near line center; the correction factor S_α accounts for these complications. As we will see below, strong Wouthuysen-Field coupling is relatively easy to achieve in practice.

The Lyα coupling efficiency also depends on the effective temperature T_c of the UV radiation field, defined in equation (31) and determined by the shape of the photon spectrum at the Lyα resonance. That the effective temperature of the radiation field *must* matter is easy to see: the energy defect between the different hyperfine splittings of the Lyα transition implies that the mixing process is sensitive to the gradient of the background spectrum near the Lyα resonance. More precisely, the procedure described near equation (34) lets us write

$$\frac{P_{01}}{P_{10}} = \frac{g_1}{g_0} \frac{n_{\rm ad} + n_{\rm ae}}{n_{\rm bd} + n_{\rm be}} \approx 3 \left(1 + \nu_0 \frac{{\rm d} \ln n_\nu}{{\rm d}\nu} \right), \tag{36}$$

where $n_\nu = c^2 J_\nu / 2\nu^2$ is the photon occupation number. Thus by comparison to equation (31), we have (neglecting stimulated emission)

$$\frac{h}{k_B T_c} = -\frac{{\rm d} \ln n_\nu}{{\rm d}\nu}. \tag{37}$$

Simple arguments show that $T_c \approx T_K$: all boil down to the observation that, so long as the medium is extremely optically thick, the enormous number of Lyα scatterings must bring the Lyα profile to a blackbody of temperature T_K near the line center. This condition is easily fulfilled in the high-redshift IGM, where the mean Lyα optical depth experienced by a photon that redshifts across the entire resonance is

$$\tau_{\rm GP} = \frac{\chi_\alpha \, n_{\rm HI}(z) \, c}{H(z) \nu_\alpha} \approx 3 \times 10^5 \, \bar{x}_{\rm HI} \left(\frac{1+z}{7} \right)^{3/2}. \tag{38}$$

The many atomic recoils during this scattering tilt the spectrum to the red and establish this equilibrium. In the limit $T_c \to T_K$ (a reasonable approximation in most situations of interest), equation (29) may be written as

$$1 - \frac{T_\gamma}{T_S} = \frac{x_c + x_\alpha}{1 + x_c + x_\alpha} \left(1 - \frac{T_\gamma}{T_K} \right). \tag{39}$$

Finally, we note that photons absorbed into higher Lyman-series transitions can also mix the hyperfine levels. However, they typically scatter only a few times before cascading to lower-energy lines; about one-third end up as Lyα photons, which then scatter just as above.

6.2.2 The 21 cm background

Note: much of the introductory physics will actually be included in earlier sections (in somewhat more detail), but is collected here for convenience as an excerpt.

Now that we have reviewed the underlying physics of the 21 cm transition, we will consider how the average 21 cm background from the high-z IGM evolves through time. We begin by examining the cosmic dark ages, when the physics is rather simple. The first step is to compute how T_K evolves with time. Energy conservation in the expanding IGM demands

$$\frac{\mathrm{d}T_K}{\mathrm{d}t} = -2H(z)T_K + \frac{2}{3}\sum_i \frac{\epsilon_i}{k_B n}, \tag{40}$$

where the first term on the right hand side is the $p\mathrm{d}V$ work from expansion and ϵ_i is the energy injected into the gas per second per unit (physical) volume through process i. Before the first nonlinear objects appear, the only relevant heating mechanism is Compton scattering between CMB photons and residual free electrons in the IGM. The heating rate from this process can be calculated from the drag force exerted by the isotropic CMB radiation field on a thermal distribution of free electrons,

$$\frac{2}{3}\frac{\epsilon_{\rm comp}}{k_B n} = \frac{\bar{x}_i}{1+f_{\rm He}+\bar{x}_i}\frac{(T_\gamma - T_K)}{t_\gamma}, \tag{41}$$

where $t_\gamma \equiv (3m_e c)/(8\sigma_T u_\gamma)$ is the Compton cooling time, $u_\gamma \propto T_\gamma^4$ is the energy density of the CMB, $f_{\rm He}$ is the helium fraction (by number), and σ_T is the Thomson cross section. The first factor on the right hand side accounts for the distribution of the energy over all free particles. Compton heating drives $T_K \to T_\gamma$ when u_γ and $\bar{x}_i$ are large; thus at sufficiently high redshifts the gas can cool no faster than the CMB, $T_K \propto (1+z)$.

Eventually, however, the gas does thermally decouple from the CMB. The recombination rate is $\dot{n}_e = -\alpha_B \bar{x}_i^2 n_b^2$, where $\alpha_B \propto T_K^{-0.7}$ is the case-B recombination coefficient. The fractional change in $\bar{x}_i$ per Hubble time is therefore

$$\frac{1}{H(z)n_e}\frac{\mathrm{d}n_e}{\mathrm{d}t} \approx 100\bar{x}_i(1+z)^{0.8}\frac{\Omega_b h^2}{\sqrt{\Omega_0 h^2}}, \tag{42}$$

where we have assumed that $T_K \propto (1+z)$ (i.e., coupling to the CMB is still strong). Freeze-out occurs when this is of order unity; at later times $\bar{x}_i$ remains roughly constant because the recombination time then exceeds the expansion timescale. Detailed numerical calculations yield $\bar{x}_i = 3.1\times10^{-4}$ after freeze-out; inserting this into equation (41), we find

$$\frac{1}{H(z)T_K}\frac{\mathrm{d}T_K}{\mathrm{d}t} \sim \frac{10^{-7}}{\Omega_b h^2}\left(T_\gamma/T_K - 1\right)(1+z)^{5/2}. \tag{43}$$

Thus thermal decoupling occurs when

$$1 + z_{\rm dec} \approx 150\left(\Omega_b h^2/0.023\right)^{2/5}. \tag{44}$$

Figure 3a shows a more detailed calculation. Compton heating begins to become inefficient at $z \sim 300$ and is negligible by $z \sim 150$. Past this point, $T_K \propto (1+z)^2$, as expected for an adiabatically expanding non-relativistic gas.

We must next determine the spin temperature. Barring any exotic processes $x_\alpha = 0$ during this epoch. But at sufficiently high redshifts, the Universe was dense enough for collisions with neutral atoms to be efficient in the mean density IGM. The effectiveness of collisional coupling can be computed exactly for any given temperature history from the rate coefficients described earlier; a convenient estimate of their importance is the critical overdensity, $\delta_{\rm coll}$, at which $x_c = 1$:

$$1 + \delta_{\rm coll} = 1.06\left[\frac{\kappa_{10}(88\ \mathrm{K})}{\kappa_{10}(T_K)}\right]\left(\frac{0.023}{\Omega_b h^2}\right)\left(\frac{70}{1+z}\right)^2, \tag{45}$$

where we have inserted the expected temperature at $1+z = 70$. Thus, in the standard history, for redshifts $z < 70$, $T_S \to T_\gamma$; by $z \sim 30$ the IGM essentially becomes invisible (see Figure 3b) The signal peaks (in absorption) at $z \sim 80$, where T_K is small but collisional coupling still efficient. Because of the simple physics involved in Figure 3, the 21 cm line offers a sensitive probe of the dark ages, at least in principle.

Figure 3b also shows that the z<30 Universe would remain invisible without luminous sources. The properties of the first galaxies will therefore determine the observability of the 21 cm background during this later era. There are two principal components to this: the thermal evolution – which is likely dominated by X-ray heating – and the UV background, which makes the 21 cm line visible against the CMB. Other heating channels, especially shocks, may also contribute but are more difficult to quantify.

Because they have relatively long mean free paths, X-rays from galaxies and quasars are likely to be the most important heating agent for the low-density IGM. In particular, photons with $E > 1.5\bar{x}_{\rm HI}^{1/3}[(1+z)/10]^{1/2}$ keV have mean free paths exceeding the Hubble length. Given our enormous uncertainties about

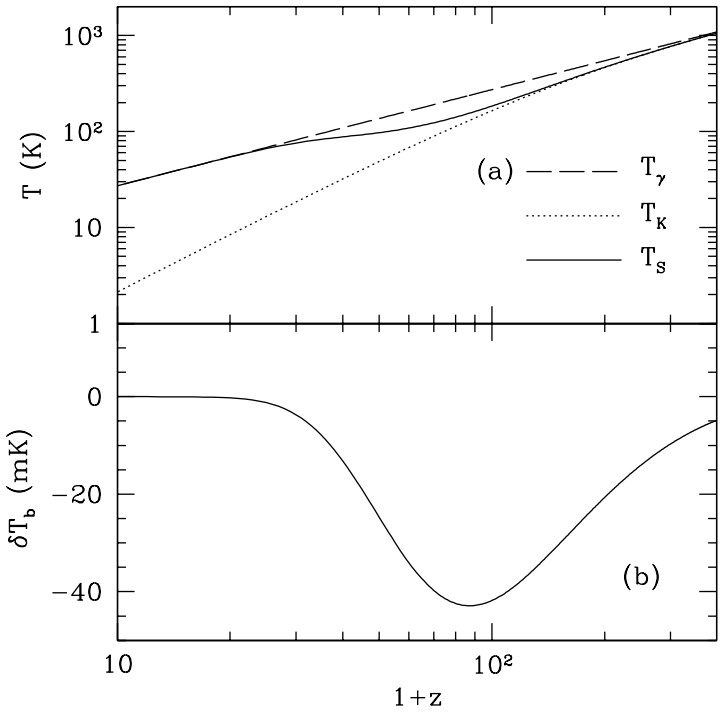

Figure 3: *(a)*: IGM temperature evolution if only adiabatic cooling and Compton heating are involved. The spin temperature T_S includes only collisional coupling. *(b)*: Differential brightness temperature against the CMB for T_S shown in panel *a*.

the nature of high-redshift objects it is of course impossible to describe the high-redshift X-ray background with any confidence. The most conservative assumption is that the local correlation between the star formation rate (SFR) and the X-ray luminosity (from 0.2–10 keV) can be extrapolated to high redshift; then

$$L_X = 3.4 \times 10^{40} f_X \left(\frac{\mathrm{SFR}}{1 \ \mathrm{M_\odot \ yr^{-1}}} \right) \ \mathrm{erg \ s^{-1}}, \tag{46}$$

where f_X is an unknown renormalization factor at high redshift. Note that the numerical factor depends on the photon energy range assumed to contribute to IGM heating; soft photons probably carry most of the energy, but they do not penetrate far into the IGM.

We can only speculate as to the accuracy of this correlation at higher redshifts. Certainly the scaling is appropriate so long as stars dominate, but f_X will likely evolve with redshift. The X-ray emission has two major sources. The first is inverse-Compton scattering off of relativistic electrons accelerated in supernovae. In the nearby Universe, only powerful starbursts have strong enough radiation fields for this to be significant; however, at high-redshifts it probably plays an increasingly important role because $u_\gamma \propto (1+z)^4$. Assuming that $\sim 5\%$ of the supernova energy is released in this form, with $\sim 10^{51}$ ergs in supernovae per 100 $\mathrm{M_\odot}$ in star formation, yields $f_X \sim 0.5$.

The second class of sources, which dominate in locally observed galaxies, are high-mass X-ray binaries, in which material from a massive main sequence star accretes onto a compact neighbor. Such systems are born as soon as the first massive stars die, only a few million years after star formation commences. So they certainly ought to exist in high-redshift galaxies, although their abundance depends on the metallicity and stellar initial mass function, and it could be especially large if very massive Population III stars dominate.

X-rays heat the IGM gas by first photoionizing a hydrogen or helium atom. The hot "primary" electron then distributes its energy through three main channels: (1) collisional ionizations, producing more secondary electrons, (2) collisional excitations of HeI (which produce photons capable of ionizing HI) and HI (which produces a Lyα background), and (3) Coulomb collisions with free electrons. The relative cross sections of these processes determine what fraction of the X-ray energy goes to heating ($f_{X,h}$), ionization ($f_{X,\mathrm{ion}}$), and excitation ($f_{X,\mathrm{coll}}$); clearly it depends on both $\bar{x}_i$ and the input photon energy. A crude approximation to these rates is:

$$\begin{aligned} f_{X,h} &\sim (1 + 2\bar{x}_i)/3 \\ f_{X,\mathrm{ion}} \sim f_{X,\mathrm{coll}} &\sim (1 - \bar{x}_i)/3. \end{aligned} \tag{47}$$

In highly ionized gas, collisions with free electrons dominate and $f_{X,h} \to 1$; in the opposite limit, the energy is split roughly equally between these processes.

Finally, to relate the X-ray emissivity to the global SFR we will assume (again for simplicity) that the SFR is proportional to the rate at which gas collapses onto virialized halos, $\mathrm{d}f_{\rm coll}/\mathrm{d}t$ (where $f_{\rm coll}$ is the fraction of matter inside of star-forming halos). In that case, we can write

$$\frac{2}{3}\frac{\epsilon_X}{k_B n H(z)} = 10^3 \text{ K } f_X \left(\frac{f_\star}{0.1}\frac{f_{X,h}}{0.2}\frac{\mathrm{d}f_{\rm coll}/\mathrm{d}z}{0.01}\frac{1+z}{10}\right), \tag{48}$$

where $f_\star$ is the star formation efficiency. It is immediately obvious that X-ray heating is quite rapid.

Of course, even if equation (46) is accurate, there may be other contributions to the X-ray background. Quasars are one obvious example, although their relevance is far from clear. The known population of bright quasars, extrapolated to higher redshifts, causes negligible heating. But these bright quasars may be only the tip of the iceberg: "miniquasars," which are rapidly accreting intermediate-mass black holes that may form from the remnants of Pop III stars, can strongly affect the thermal history. For example, if the "Magorrian relation" between black hole and stellar mass holds at high redshifts, the equivalent normalization factor would be $f_X \sim 10$.

With the thermal evolution in hand, we now turn to the spin temperature. Recall from equation (45) that collisions are inefficient at $z<30$, so we must rely on the Wouthuysen-Field effect. Of course (as for the X-ray background), we cannot yet predict the detailed evolution of J_α, because it depends on the star formation history as well as any other radiation background (quasars, etc.). But we can make an educated guess by assuming that it traces the star formation rate, which is again (roughly) proportional to the rate at which matter collapses into galaxies. We therefore write the comoving emissivity at frequency ν as

$$\epsilon(\nu, z) = f_\star\, \bar{n}_b^c\, \epsilon_b(\nu)\, \frac{\mathrm{d}f_{\rm coll}}{\mathrm{d}t}, \tag{49}$$

where $\bar{n}_b^c$ is the comoving number density of baryons and $\epsilon_b(\nu)$ is the number of photons produced in the frequency interval $\nu \pm \mathrm{d}\nu/2$ per baryon incorporated into stars. Here we are only interested in photons between Lyα and the Lyman-limit. Although real spectra are rather complicated, a useful quantity is the total number N_α of photons per baryon in this interval. For low-metallicity Pop II stars and very massive Pop III stars, this is $N_\alpha = 9690$ and $N_\alpha = 4800$, respectively.

Although only Lyα photons efficiently couple to T_S, higher Lyman-series photons contribute by cascading to Lyα. The average background at ν_α is

$$\begin{aligned} J_\alpha(z) &= \sum_{n=2}^{n_{\rm max}} J_\alpha^{(n)}(z) \\ &= \sum_{n=2}^{n_{\rm max}} f_{\rm rec}(n) \int_z^{z_{\rm max}(n)} \mathrm{d}z' \, \frac{(1+z)^2}{4\pi}\frac{c}{H(z')}\,\epsilon(\nu_n', z'), \end{aligned} \tag{50}$$

where ν_n' is the frequency at redshift z' that redshifts into the Lyn resonance at redshift z, $z_{\rm max}(n)$ is the largest redshift from which a photon can redshift into the Lyn resonance, and $f_{\rm rec}(n)$ is the fraction of Lyn photons that actually cascade through Lyα and induce strong coupling ($f_{\rm rec} \sim 1/3$ for higher-series transitions). Once we know J_α, we can compute the coupling coefficient x_α from equation (35).

In detail, we must also include other processes that produce Lyα photons. For example, collisional excitation by X-rays can be important (see eq. 47). The coupling coefficient induced by these line photons is

$$x_\alpha^{\rm X-ray} \sim 0.05\, S_\alpha\, f_X \left(\frac{f_{X,\rm coll}}{1/3}\frac{f_\star}{0.1}\frac{\mathrm{d}f_{\rm coll}/\mathrm{d}z}{0.01}\right)\left(\frac{1+z}{10}\right)^3. \tag{51}$$

Here we have substituted the same emissivity as in equation (48); thus heating is accompanied by a small, though far from negligible, Lyα background. This process is particularly important near star-forming galaxies, where most soft X-rays are absorbed.

The next step is of course reionization itself. For a very simple model, we will make the usual assumption that ionizing photons are produced inside of galaxies, so that their production rate can be associated with the star formation rate in a similar way to the Lyα radiation background and our X-ray heating model (see eqs. 48 and 49). In the most basic approximation, we simply assign a fixed average ionizing efficiency across all galaxies, so that

$$\bar{x}_i = \zeta f_{\rm coll}/(1+\bar{n}_{\rm rec}), \tag{52}$$

where $\bar{n}_{\rm rec}$ is the mean number of recombinations per ionized hydrogen atom and the ionizing efficiency is

$$\zeta = A_{\rm He} f_\star f_{\rm esc} N_{\rm ion}. \tag{53}$$

In this expression, $f_{\rm esc}$ is the fraction of ionizing photons that escape their host galaxy into the IGM, $N_{\rm ion}$ is the mean number of ionizing photons produced per stellar baryon, and $A_{\rm He} = 4/(4-3Y_p) = 1.22$, where Y_p is the mass fraction of helium, is a correction factor to convert the number of ionizing photons per baryon in stars to the fraction of ionized hydrogen.[3]

We can use local measurements of $f_\star$, $f_{\rm esc}$, and $N_{\rm ion}$ to guide our choices for these parameters, though the extrapolation to high redshifts is always difficult. Efficiencies $f_\star \sim 10\%$ are reasonable for the local Universe, but so little gas has collapsed by $z = 6$ that this does not directly constrain the high-redshift value. Appropriate values for Pop III stars are even more uncertain. To the extent that each halo can form only a single very massive (> 100 M$_\odot$) star that enriches the entire halo, $f_\star \sim (\Omega_m/\Omega_b)M_\star/M_h < 10^{-3}$. The UV escape fraction is small in nearby star-forming galaxies (including the Milky Way), with many upper limits $f_{\rm esc} < 6\%$ and only a few positive detections (at comparable levels). Theoretically the problem is equally difficult, because it depends on the spatial distribution of hot stars and absorbing gas in the ISM. Some studies suggest that the escape fraction in high-redshift galaxies could be much higher than the detected values (generally because higher specific star formation rates allow supernovae to blow transparent windows through the ISM), but others predict that it will remain small. $N_{\rm ion}$ depends only on the stellar initial mass function and metallicity. Convenient approximations are $N_{\rm ion} \approx 4000$ for $Z = 0.05$ Z$_\odot$ Pop II stars with a Scalo IMF and $N_{\rm ion} \approx 40,000$ for very massive Pop III stars. Note that the latter assumes that *all* Pop III stars are massive; metal-free stars with a normal Salpeter IMF are only ~ 1.6 times more efficient than their Pop II counterparts.

The most basic information contained in the 21 cm background is the sequence in which these various backgrounds become important, and this information can be quantified simply with a few critical points. There are five such points in the 21 cm history that divide the signal into several distinct epochs. The first is $z_{\rm dec}$, when Compton heating becomes inefficient and $T_K < T_\gamma$ for the first time (eq. 44). This marks the earliest epoch for which 21 cm observations are possible even in principle. The second transition is when the density falls below $\delta_{\rm coll}$ (see eq. 45), at which point $T_S \to T_\gamma$ and the IGM signal vanishes. These two points are well-specified by atomic physics processes.

The remaining transition points are determined by luminous sources, so their timing is much more uncertain. These are the redshift z_h at which the IGM is heated above T_γ, the redshift z_c at which $x_\alpha = 1$ so that the Wouthuysen-Field mechanism couples T_S and T_K, and the redshift of reionization z_r. We first ask whether Lyα coupling precedes the other two transitions. The net X-ray heat input ΔT_c at z_c is

$$\frac{\Delta T_c}{T_\gamma} \sim 0.08 f_X \left(\frac{f_{X,h}}{0.2}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{9690}{N_\alpha}\frac{1}{S_\alpha}\frac{0.023}{\Omega_b h^2}\right)\left(\frac{20}{1+z}\right)^3, \tag{54}$$

where $\Delta f_{\rm coll} \sim f_{\rm coll}$ is the effective collapse fraction appearing in the integrals of equation (50). Note that ΔT_c is independent of $f_\star$ because both the coupling and heating rates are proportional to the star formation rate. Interestingly, for our fiducial (Pop II) parameters z_c precedes z_h. This could create a significant absorption epoch whose properties offer a meaningful probe of the first sources. For example, very massive Pop III stars have a smaller N_α, and an early miniquasar population could completely eliminate the absorption epoch.

A similar estimate of the ionization fraction $\bar{x}_{i,c}$ at z_c yields

$$\bar{x}_{i,c} \sim 0.05 \left(\frac{f_{\rm esc}}{1+\bar{n}_{\rm rec}}\frac{N_{\rm ion}}{N_\alpha}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{1}{S_\alpha}\frac{0.023}{\Omega_b h^2}\right)\left(\frac{20}{1+z}\right)^2. \tag{55}$$

For Pop II stars, $N_{\rm ion}/N_\alpha \approx 0.4$; thus even in the worst case of $f_{\rm esc} = 1$ and $\bar{n}_{\rm rec} = 0$ coupling would become efficient during the initial stages of reionization. However, very massive Pop III stars have much harder spectra, with $N_{\rm ion}/N_\alpha \approx 7$. In principle, it is therefore possible for Pop III stars to reionize the universe *before* z_c, although that would require *extremely* unusual parameters. Such histories cannot be ruled out at present, but we regard them as exceedingly unlikely. Histories with $\bar{x}_{i,c} \ll 1$ are much more plausible, at least given our theoretical prejudices about high-redshift sources.

Finally, we ask whether the IGM will appear in absorption or emission during reionization. Combining equations (48) and (52), we have

$$\frac{\Delta T}{T_\gamma} \sim \left(\frac{\bar{x}_i}{0.025}\right)\left(f_X \frac{f_{X,h}}{f_{\rm esc}}\frac{4800}{N_{\rm ion}}\frac{10}{1+z}\right)(1+\bar{n}_{\rm rec}) \tag{56}$$

for the heat input ΔT as a function of $\bar{x}_i$. Thus, provided $f_X > 1$, the IGM will be much warmer than the CMB during the bulk of reionization. This is convenient in that δT_b becomes independent of T_S when $T_S \gg T_\gamma$, so it is easier to isolate the effects of the ionization field. Significant absorption during reionization

[3]Here we have assumed that helium is singly ionized along with hydrogen, because their ionization potentials are relatively close.

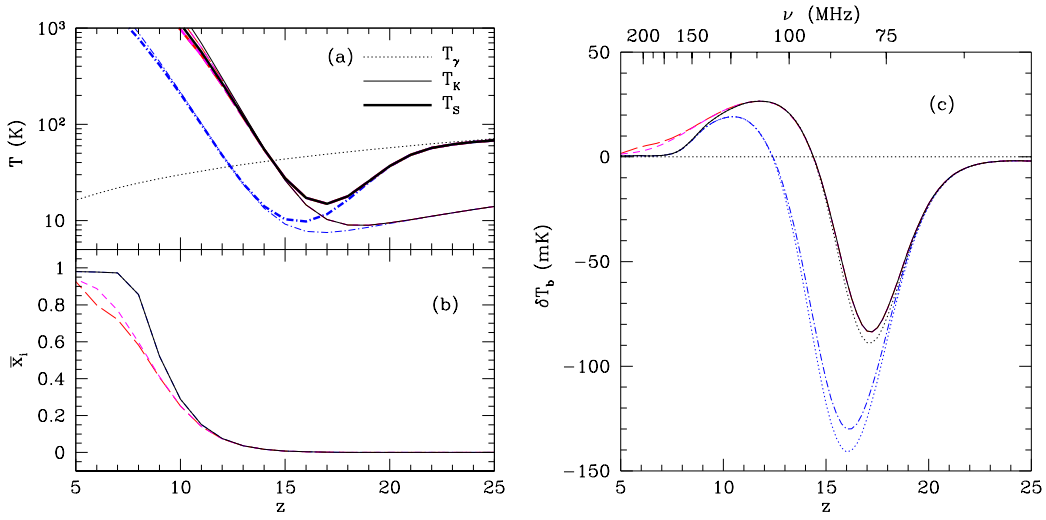

Figure 4: Global IGM histories for Pop II stars. The solid curves take our fiducial parameters without feedback. The dot-dashed curve takes $f_X = 0.2$. The short- and long-dashed curves include strong photoheating feedback. *(a)*: Thermal properties. *(b)*: Ionized fraction. *(c)*: Differential brightness temperature against the CMB. In this panel, the two dotted lines show δT_b without including shock heating.

becomes more plausible for very massive Pop III stars, because they have much larger ionizing efficiencies (although their remnants may also induce correspondingly large X-ray heating).

We will now use some representative models to illustrate these qualitative features in a more concrete fashion. We begin with a fiducial set of Pop II parameters. We ignore feedback (of all kinds) and take $m_{\rm min}$ to correspond to $T_{\rm vir} = 10^4$ K, $f_\star = 0.1$, $f_{\rm esc} = 0.1$, $f_X = 1$, $N_{\rm ion} = 4000$, and $N_\alpha = 9690$. (Thus $\zeta = 40$ for this model.) Figure 4*a* shows the resulting temperature history. The dotted curve is T_γ, the thin solid curve is T_K, and the thick solid curve is T_S. As expected from equation (54), in this case we do indeed find that $z_c > z_h$; specifically, $z_c \approx 18$ and $z_h \approx 14$. Clearly Lyα coupling is extremely efficient for normal stars. The solid curve in Figure 4*b* shows the corresponding ionization history. It increases smoothly and rapidly over a redshift interval of $\Delta z \sim 5$, ending at $z_r \sim 7$. That is of course purely a function of our choice for ζ, but other values do not strongly affect the width.

Figure 4*c* shows the corresponding 21 cm brightness temperature decrement δT_b relative to the CMB. Here we have also labeled the corresponding (observed) frequency ν for convenience. The signal clearly has interesting structure. At the highest frequencies, reionization causes a steady decline in the signal, with $|{\rm d}\delta\bar{T}_b/{\rm d}\nu| \sim 1$ mK MHz^{-1}. In this model, recombinations are relatively inefficient; the only way to significantly increase the gradient during reionization would be with some positive feedback mechanism. However, as illustrated by the dashed curves, it is relatively easy to slow reionization. These curves use two models for photoheating feedback in which the minimum virial temperature for galaxy formation increases to $T_h = 2 \times 10^5$ K in photoheated regions.

Figure 4*c* contains an even more striking feature at higher redshifts. At $z \sim 30$, the IGM is nearly invisible even though $T_K \ll T_\gamma$ (see Fig. 3). However, as the first galaxies form, the Wouthuysen-Field effect drives $T_S \rightarrow T_K$. Because $z_c > z_h$, this produces a relatively strong absorption signal ($\delta T_b \approx -80$ mK) over the range $z \sim 21$–14 (or $\nu \sim 70$–95 MHz). However, the IGM still heats up well before reionization begins in earnest, making $\delta\bar{T}_b$ nearly independent of T_S throughout reionization.

Figure 5 shows similar histories for very massive Pop III stars. The solid curves take $m_{\rm min}$ to correspond to $T_{\rm vir} = 10^4$ K, $f_\star = 0.01$, $f_{\rm esc} = 0.1$, $f_X = 1$, $N_{\rm ion} = 30,000$, and $N_\alpha = 4800$, yielding $\zeta = 30$. Although the thermal history is qualitatively similar to the Pop II case, it has $z_c \sim 13$ and $z_h \sim 11$. Thus the absorption epoch is somewhat narrower, and it is also weaker because Pop III stars produce relatively few Lyα photons. As a result, T_S does not approach T_K until the IGM is already hot. Thus, if very massive Pop III stars dominate, the absorption epoch will be considerably weaker, with gradients about half as large as the Pop II case. Moreover, z_h is relatively close to z_r, so T_S does not saturate until after reionization begins. It may therefore be somewhat difficult to separate T_S and x_i at the beginning of reionization.

Obviously, measuring this background could offer strong constraints on high-redshift star formation. The other curves in Figures 4–5 illustrate the range of features we expect. However, they all share one crucial property: z_c occurs long before reionization, so we can safely expect *some* signal from the high-redshift IGM.

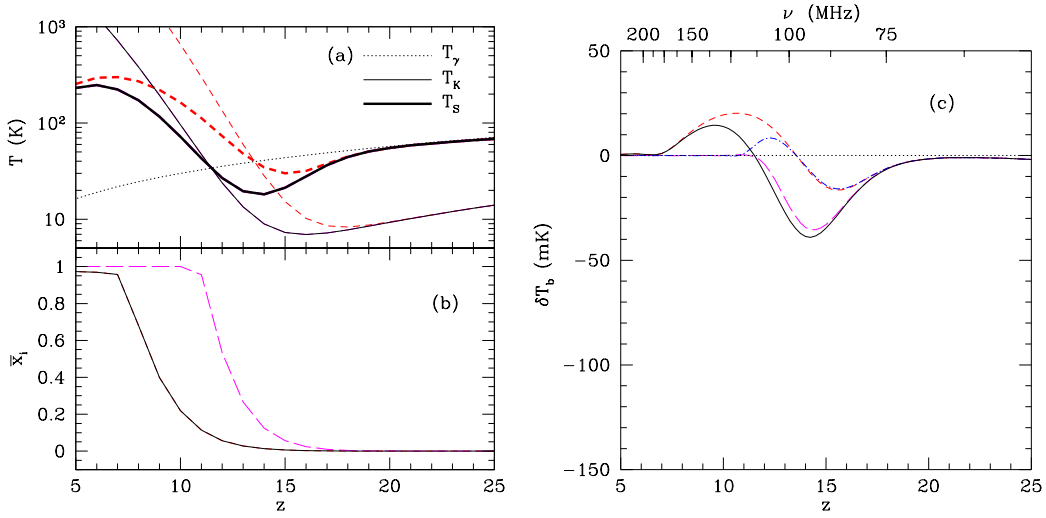

Figure 5: Global IGM histories for very massive Pop III stars. Panels are the same as in Fig. 4. The solid curve takes our fiducial Pop III parameters. The long-dashed lines take $f_{\rm esc} = 1$, the short-dashed lines take $f_X = 5$, and the dot-dashed line (shown only in c) assumes $f_{\rm esc} = 1$ and $f_X = 5$.