The First Galaxies

Abraham Loeb and Steven R. Furlanetto

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

To our families

Contents

Preface

This book captures the latest exciting developments concerning one of the unsolved mysteries about our origins: *how did the first stars and galaxies form?* Most research on this question has been theoretical so far. But the next few years will bring about a new generation of large telescopes with unprecedented sensitivity that promise to supply a flood of data about the infant Universe during its first billion years after the Big Bang. Among the new observatories are the James Webb Space Telescope (JWST) – the successor to the Hubble Space Telescope, and three extremely large telescopes on the ground (ranging from 24 to 42 meters in diameter), as well as several new arrays of dipole antennae operating at low radio frequencies. The fresh data on the first galaxies and the diffuse gas in between them will test existing theoretical ideas about the formation and radiative effects of the first galaxies, and might even reveal new physics that has not yet been anticipated. This emerging interface between theory and observation will constitute an ideal opportunity for students considering a research career in astrophysics or cosmology. With this in mind, the book is intended to provide a self-contained introduction to research on the first galaxies at a technical level appropriate for a graduate student.

Various introductory sections of this book are based on an undergraduate-level book, entitled "How Did the First Stars and Galaxies Form?" by one of us (A.L.), which followed a cosmology class that he had taught over the past decade in the Astronomy and Physics departments at Harvard University. Other parts relate to overviews that both of us wrote over the past decade in the form of review articles. Where necessary, selected references are given to advanced papers and other review articles in the scientific literature.

The writing of this book was made possible thanks to the help we received from a large number of individuals. First and foremost, ... Special thanks go to ... for their careful reading of the book and detailed comments. We also thank Joey Munoz and Ramesh Narayan for their help with two plots. Finally, we are particularly grateful to our families for their support and patience during our lengthy pregnancy period with the book.

–A. L. & S. F.

Chapter One

Introduction

1.1 PRELIMINARY REMARKS

As the Universe expands, galaxies get separated from one another, and the average density of matter over a large volume of space is reduced. If we imagine playing the cosmic movie in reverse and tracing this evolution backwards in time, we would infer that there must have been an instant when the density of matter was infinite. This moment in time is the " Big Bang", before which we cannot reliably extrapolate our history. But even before we get all the way back to the Big Bang, there must have been a time when stars like our Sun and galaxies like our Milky Way[i] did not exist, because the Universe was denser than they are. If so, *how and when did the first stars and galaxies form?*

Primitive versions of this question were considered by humans for thousands of years, long before it was realized that the Universe expands. Religious and philosophical texts attempted to provide a sketch of the big picture from which people could derive the answer. In retrospect, these attempts appear heroic in view of the scarcity of scientific data about the Universe prior to the twentieth century. To appreciate the progress made over the past century, consider, for example, the biblical story of Genesis. The opening chapter of the Bible asserts the following sequence of events: first, the Universe was created, then light was separated from darkness, water was separated from the sky, continents were separated from water, vegetation appeared spontaneously, stars formed, life emerged, and finally humans appeared on the scene.[ii] Instead, the modern scientific order of events begins with the Big Bang, followed by an early period in which light (radiation) dominated and then a longer period dominated by matter, leading to the appearance of stars, planets, life on Earth, and eventually humans. Interestingly, the starting and end points of both versions are the same.

Cosmology is by now a mature empirical science. We are privileged to live in a time when the story of genesis (how the Universe started and developed) can be critically explored by direct observations. Because of the finite time it takes light to travel to us from distant sources, we can see images of the Universe when it was younger by looking deep into space through powerful telescopes.

Existing data sets include an image of the Universe when it was 400 thousand

[i] A **star** is a dense, hot ball of gas held together by gravity and powered by nuclear fusion reactions. A **galaxy** consists of a luminous core made of stars or cold gas surrounded by an extended halo of *dark matter*.

[ii] Of course, it is possible to interpret the biblical text in many possible ways. Here I focus on a plain reading of the original Hebrew text.

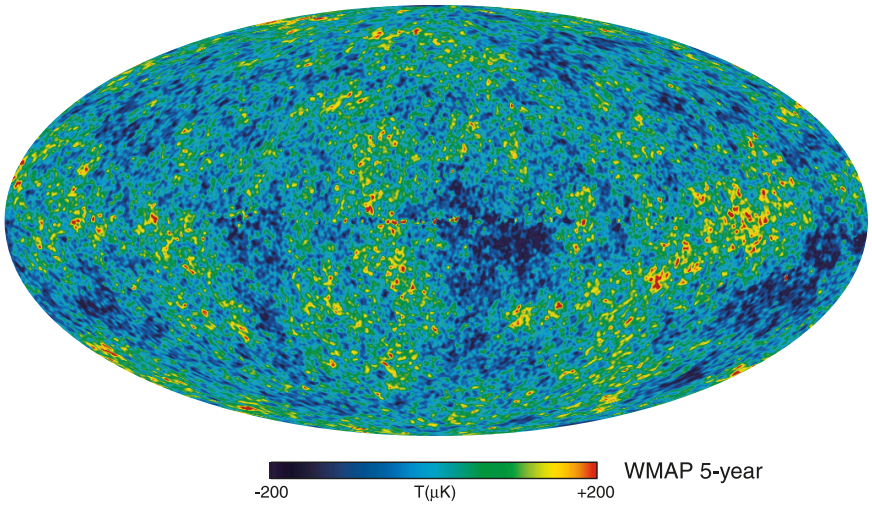

Figure 1.1 Image of the Universe when it first became transparent, 400 thousand years after the Big Bang, taken over five years by the *Wilkinson Microwave Anisotropy Probe* (WMAP) satellite (http://map.gsfc.nasa.gov/). Slight density inhomogeneities at the level of one part in $\sim 10^5$ in the otherwise uniform early Universe imprinted hot and cold spots in the temperature map of the cosmic microwave background on the sky. The fluctuations are shown in units of μK, with the unperturbed temperature being 2.73 K. The same primordial inhomogeneities seeded the large-scale structure in the present-day Universe. The existence of background anisotropies was predicted in a number of theoretical papers three decades before the technology for taking this image became available.

years old (in the form of the cosmic microwave background in Figure 1.1), as well as images of individual galaxies when the Universe was older than a billion years. But there is a serious challenge: in between these two epochs was a period when the Universe was dark, stars had not yet formed, and the cosmic microwave background no longer traced the distribution of matter. And this is precisely the most interesting period, when the primordial soup evolved into the rich zoo of objects we now see. *How can astronomers see this dark yet crucial time?*

The situation is similar to having a photo album of a person that begins with the first ultra-sound image of him or her as an unborn baby and then skips to some additional photos of his or her years as teenager and adult. The late photos do not simply show a scaled up version of the first image. We are currently searching for the missing pages of the cosmic photo album that will tell us how the Universe evolved during its infancy to eventually make galaxies like our own Milky Way.

The observers are moving ahead along several fronts. The first involves the construction of large infrared telescopes on the ground and in space that will provide us with new (although rather expensive!) photos of galaxies in the Universe at intermediate ages. Current plans include ground-based telescopes which are 24-42 meter in diameter, and NASA's successor to the Hubble Space Telescope, the James Webb Space Telescope. In addition, several observational groups around the globe are constructing radio arrays that will be capable of mapping the three-dimensional distribution of cosmic hydrogen left over from the Big Bang in the infant Universe. These arrays are aiming to detect the long-wavelength (redshifted 21-cm) radio emission from hydrogen atoms. Coincidentally, this long wavelength (or low frequency) overlaps with the band used for radio and television broadcasting, and so these telescopes include arrays of regular radio antennas that one can find in electronics stores. These antennas will reveal how the clumpy distribution of neutral hydrogen evolved with cosmic time. By the time the Universe was a few hundreds of millions of years old, the hydrogen distribution had been punched with holes like swiss cheese. These holes were created by the ultraviolet radiation from the first galaxies and black holes, which ionized the cosmic hydrogen in their vicinity.

Theoretical research has focused in recent years on predicting the signals expected from the above instruments and on providing motivation for these ambitious observational projects.

1.2 STANDARD COSMOLOGICAL MODEL

1.2.1 Cosmic Perspective

In 1915 Einstein came up with the general theory of relativity. He was inspired by the fact that all objects follow the same trajectories under the influence of gravity (the so-called "equivalence principle," which by now has been tested to better than one part in a trillion), and realized that this would be a natural result if space-time is curved under the influence of matter. He wrote down an equation describing how the distribution of matter (on one side of his equation) determines the curvature of space-time (on the other side of his equation). He then applied his equation to

describe the global dynamics of the Universe.

Back in 1915 there were no computers available, and Einstein's equations for the Universe were particularly difficult to solve in the most general case. It was therefore necessary for Einstein to alleviate this difficulty by considering the simplest possible Universe, one that is homogeneous and isotropic. Homogeneity means uniform conditions everywhere (at any given time), and isotropy means the same conditions in all directions when looking out from one vantage point. The combination of these two simplifying assumptions is known as the *cosmological principle*.

The universe can be homogeneous but not isotropic: for example, the expansion rate could vary with direction. It can also be isotropic and not homogeneous: for example, we could be at the center of a spherically-symmetric mass distribution. But if it is isotropic around *every* point, then it must also be homogeneous. Isotropy is well established for the distribution of faint radio sources, optical 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 and significantly inhomogeneous in spherical shells around our special location, is also excluded based on surveys of galaxies and quasars.

Under the simplifying assumptions associated with the cosmological principle, Einstein and his contemporaries were able to solve the equations. They were looking for their "lost keys" (solutions) under the "lamppost" (simplifying assumptions), but the real Universe is not bound by any contract to be the simplest that we can imagine. In fact, it is truly remarkable in the first place that we dare describe the conditions across vast regions of space based on the blueprint of the laws of physics that describe the conditions here on Earth. Our daily life teaches us too often that we fail to appreciate complexity, and that an elegant model for reality is often too idealized for describing the truth (along the lines of approximating a cow as a spherical object).

Back in 1915 Einstein had the wrong notion of the Universe; at the time people associated the Universe with the Milky Way galaxy and regarded all the "nebulae," which we now know are distant galaxies, as constituents within our own Milky Way galaxy. Because the Milky Way is not expanding, Einstein attempted to reproduce a static universe with his equations. This turned out to be possible after adding a cosmological constant, whose negative gravity would exactly counteract that of matter. However, later Einstein realized that this solution is unstable: a slight enhancement in density would make the density grow even further. As it turns out, there are no stable static solution to Einstein's equations for a homogeneous and isotropic Universe. The Universe must either be expanding or contracting. Less than a decade later, Edwin Hubble discovered that the nebulae previously considered to be constituents of the Milky Way galaxy are receding away from us at a speed v that is proportional to their distance r, namely $v = H_0 r$ with H_0 a spatial constant (which could evolve with time), commonly termed the *Hubble constant*.[iii] Hubble's data indicated that the Universe is expanding.

[iii] The redshift data examined by Hubble was mostly collected by Vesto Slipher a decade earlier and only partly by Hubble's assistant, Milton L. Humason. The linear local relation between redshift and distance was first formulated by Georges Lemaître in 1927, two years prior to the observational paper written by Hubble and Humason.

Einstein was remarkably successful in asserting the cosmological principle. As it turns out, our latest data indicates that the real Universe is homogeneous and isotropic on the largest observable scales to within one part in a hundred thousand. Fortuitously, Einstein's simplifying assumptions turned out to be extremely accurate in describing reality: *the keys were indeed lying next to the lamppost.* Our Universe happens to be the simplest we could have imagined, for which Einstein's equations can be easily solved.

Why was the Universe prepared to be in this special state? Cosmologists were able to go one step further and demonstrate that an early phase transition, called *cosmic inflation* – during which the expansion of the Universe accelerated exponentially, could have naturally produced the conditions postulated by the cosmological principle. One is left to wonder whether the existence of inflation is just a fortunate consequence of the fundamental laws of nature, or whether perhaps the special conditions of the specific region of space-time we inhabit were selected out of many random possibilities elsewhere by the prerequisite that they allow our existence. The opinions of cosmologists on this question are split.

1.2.2 Origin of Structure

Hubble's discovery of the expansion of the Universe has immediate implications with respect to the past and future of the Universe. If we reverse in our mind the expansion history back in time, we realize that the Universe must have been denser in its past. In fact, there must have been a point in time where the matter density was infinite, at the moment of the so-called Big Bang. Indeed we do detect relics from a hotter denser phase of the Universe in the form of light elements (such as deuterium, helium and lithium) as well as the Cosmic Microwave Background (CMB). At early times, this radiation coupled extremely well to the cosmic gas and obtained a spectrum known as blackbody, that was predicted a century ago to characterize matter and radiation in equilibrium. The CMB provides the best example of a blackbody spectrum we have.

To get a rough estimate of when the Big Bang occurred, we may simply divide the distance of all galaxies by their recession velocity. This gives a unique answer, $\sim r/v \sim 1/H_0$, which is independent of distance.[iv] The latest measurements of the Hubble constant give a value of $H_0 \approx 70$ kilometers per second per Megaparsec,[v] implying a current age for the Universe $1/H_0$ of 14 billion years (or 5×10^{17} seconds).

The second implication concerns our future. A fortunate feature of a spherically-symmetric Universe is that when considering a sphere of matter in it, we are allowed to ignore the gravitational influence of everything outside this sphere. If we empty the sphere and consider a test particle on the boundary of an empty void

[iv] Although this is an approximate estimate, it turns out to be within a few percent of the true age of our Universe owing to a coincidence. The cosmic expansion at first decelerated and then accelerated with the two almost canceling each other out at the present-time, giving the same age as if the expansion were at a constant speed (as would be strictly true only in an empty Universe).

[v] A megaparsec (abbreviated as 'Mpc') is equivalent to 3.086×10^{24} centimeter, or roughly the distance traveled by light in three million years.

embedded in a uniform Universe, the particle will experience no net gravitational acceleration. This result, known as Birkhoff's theorem, is reminiscent of Newton's "iron sphere theorem." It allows us to solve the equations of motion for matter on the boundary of the sphere through a local analysis without worrying about the rest of the Universe. Therefore, if the sphere has exactly the same conditions as the rest of the Universe, we may deduce the global expansion history of the Universe by examining its behavior. If the sphere is slightly denser than the mean, we will infer how its density contrast will evolve relative to the background Universe.

The equation describing the motion of a spherical shell of matter is identical to the equation of motion of a rocket launched from the surface of the Earth. The rocket will escape to infinity if its kinetic energy exceeds its gravitational binding energy, making its total energy positive. However, if its total energy is negative, the rocket will reach a maximum height and then fall back. In order to figure out the future evolution of the Universe, we need to examine the energy of a spherical shell of matter relative to the origin. With a uniform density ρ, a spherical shell of radius r would have a total mass $M = \rho \times \left(\frac{4\pi}{3} r^3\right)$ enclosed within it. Its energy per unit mass is the sum of the kinetic energy due to its expansion speed $v = Hr$, $\frac{1}{2}v^2$, and its potential gravitational energy, $-GM/r$ (where G is Newton's constant), namely $E = \frac{1}{2}v^2 - \frac{GM}{r}$. By substituting the above relations for v and M, it can be easily shown that $E = \frac{1}{2}v^2(1 - \Omega)$, where $\Omega = \rho/\rho_c$ and $\rho_c = 3H^2/8\pi G$ is defined as the *critical density*. We therefore find that there are three possible scenarios for the cosmic expansion. The Universe has either: **(i)** $\Omega > 1$, making it gravitationally bound with $E < 0$ – *such a "closed Universe" will turn-around and end up collapsing towards a "big crunch"*; **(ii)** $\Omega < 1$, making it gravitationally unbound with $E > 0$ – *such an "open Universe" will expand forever*; or the borderline case **(iii)** $\Omega = 1$, making the Universe marginally bound or "flat" with $E = 0$.

Einstein's equations relate the geometry of space to its matter content through the value of Ω: an open Universe has a geometry of a saddle with a negative spatial curvature, a closed Universe has the geometry of a spherical globe with a positive curvature, and a flat Universe has a flat geometry with no curvature. Our observable section of the Universe appears to be flat.

Now we are at a position to understand how objects, like the Milky Way galaxy, have formed out of small density inhomogeneities that get amplified by gravity.

Let us consider for simplicity the background of a marginally bound (flat) Universe which is dominated by matter. In such a background, only a slight enhancement in density is required for exceeding the critical density ρ_c. Because of Birkhoff's theorem, a spherical region that is denser than the mean will behave as if it is part of a closed Universe and increase its density contrast with time, while an underdense spherical region will behave as if it is part of an open Universe and appear more vacant with time relative to the background, as illustrated in Figure 1.2. Starting with slight density enhancements that bring them above the critical value ρ_c, the overdense regions will initially expand, reach a maximum radius, and then collapse upon themselves (like the trajectory of a rocket launched straight up, away from the center of the Earth). An initially slightly inhomogeneous Universe

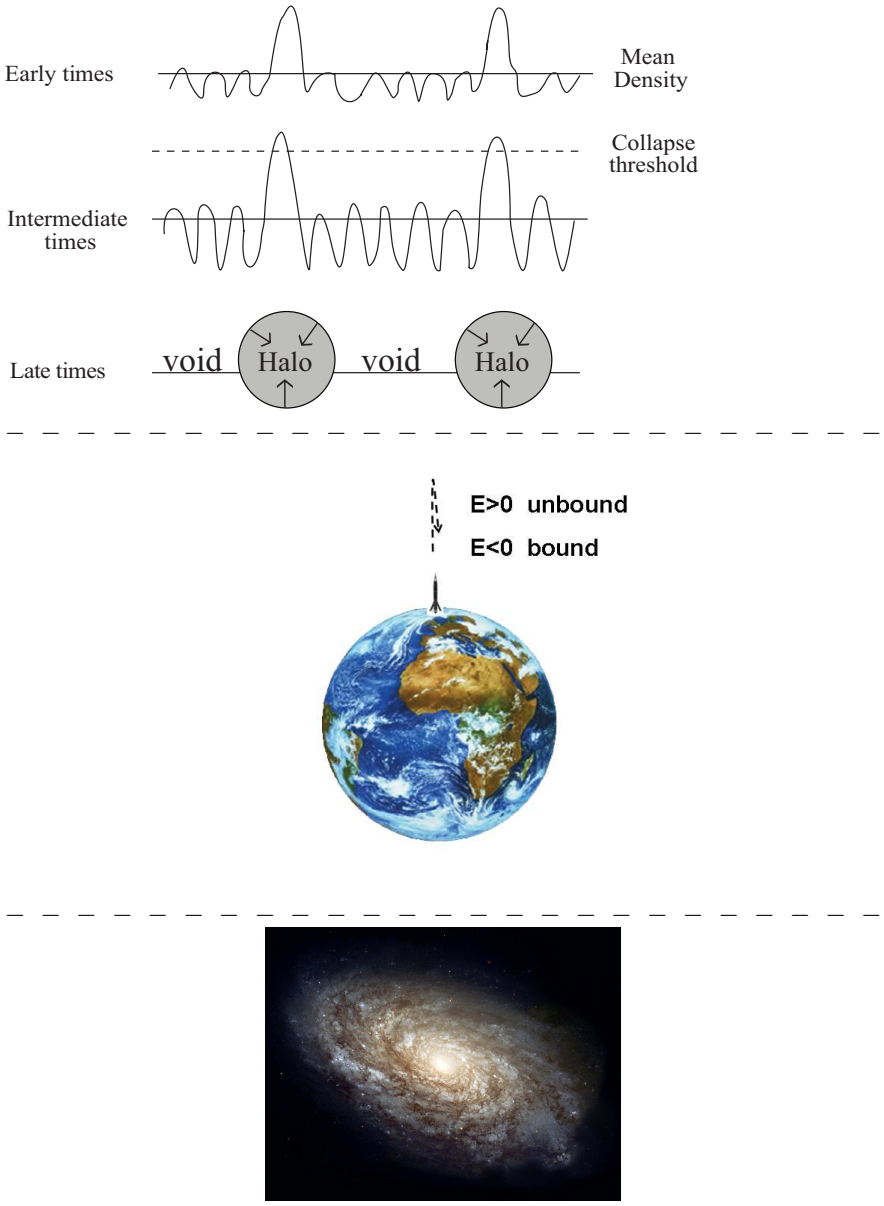

Figure 1.2 *Top:* Schematic illustration of the growth of perturbations to collapsed halos through gravitational instability. Once the overdense regions exceed a threshold density contrast above unity, they turn around and collapse to form halos. The material that makes the halos originated in the voids that separate them. *Middle:* A simple model for the collapse of a spherical region. The dynamical fate of a rocket which is launched from the surface of the Earth depends on the sign of its energy per unit mass, $E = \frac{1}{2}v^2 - GM_\oplus/r$. The behavior of a spherical shell of matter on the boundary of an overdense region (embedded in a homogeneous and isotropic Universe) can be analyzed in a similar fashion. *Bottom:* A collapsing region may end up as a galaxy, like NGC 4414, shown here (image credit: NASA and ESA). The halo gas cools and condenses to a compact disk surrounded by an extended dark matter halo.

will end up clumpy, with collapsed objects forming out of overdense regions. The material to make the objects is drained out of the intervening underdense regions, which end up as voids.

The Universe we live in started with primordial density perturbations of a fractional amplitude $\sim 10^{-5}$. The overdensities were amplified at late times (once matter dominated the cosmic mass budget) up to values close to unity and collapsed to make objects, first on small scales. We have not yet seen the first small galaxies that started the process that eventually led to the formation of big galaxies like the Milky Way. The search for the first galaxies is a search for our origins.

Life as we know it on planet Earth requires water. The water molecule includes oxygen, an element that was not made in the Big Bang and did not exist until the first stars had formed. Therefore our form of life could not have existed in the first hundred millions of years after the Big Bang, before the first stars had formed. There is also no guarantee that life will persist in the distant future.

1.2.3 Geometry of Space

How can we tell the difference between the flat surface of a book and the curved surface of a balloon? A simple way would be to draw a triangle of straight lines between three points on those surfaces and measure the sum of the three angles of the triangle. The Greek mathematician Euclid demonstrated that the sum of these angles must be 180 degrees (or π radians) on a flat surface. Twenty-one centuries later, the German mathematician Bernhard Riemann extended the field of geometry to curved spaces, which played an important role in the development of Einstein's general theory of relativity. For a triangle drawn on a positively curved surface, like that of a balloon, the sum of the angles is larger than 180 degrees. (This can be easily figured out by examining a globe and noticing that any line connecting one of the poles to the equator opens an angle of 90 degrees relative to the equator. Adding the third angle in any triangle stretched between the pole and the equator would surely result in a total of more than 180 degrees.) According to Einstein's equations, the geometry of the Universe is dictated by its matter content; in particular, the Universe is flat only if the total Ω equals unity. *Is it possible to draw a triangle across the entire Universe and measure its geometry?*

Remarkably, the answer is **yes**. At the end of the twentieth century cosmologists were able to perform this experiment[1] by adopting a simple yardstick provided by the early Universe. The familiar experience of dropping a stone in the middle of a pond results in a circular wave crest that propagates outwards. Similarly, perturbing the smooth Universe at a single point at the Big Bang would have resulted in a spherical sound wave propagating out from that point. The wave would have traveled at the speed of sound, which was of order the speed of light c (or more precisely, $\frac{1}{\sqrt{3}}c$) early on when radiation dominated the cosmic mass budget. At any given time, all the points extending to the distance traveled by the wave are affected by the original pointlike perturbation. The conditions outside this "sound horizon" will remain uncorrelated with the central point, because acoustic information has not been able to reach them at that time. The temperature fluctuations of the CMB trace the simple sum of many such pointlike perturbations that were generated in

the Big Bang. The patterns they delineate would therefore show a characteristic correlation scale, corresponding to the sound horizon at the time when the CMB was produced, 400 thousand years after the Big Bang. By measuring the apparent angular scale of this "standard ruler" on the sky, known as the acoustic peak in the CMB, and comparing it to theory, experimental cosmologists inferred from the simple geometry of triangles that the Universe is flat.

The inferred flatness is a natural consequence of the early period of vast expansion, known as cosmic inflation, during which any initial curvature was flattened. Indeed a small patch of a fixed size (representing our current observable region in the cosmological context) on the surface of a vastly inflated balloon would appear nearly flat. The sum of the angles on a non-expanding triangle placed on this patch would get arbitrarily close to 180 degrees as the balloon inflates.

1.2.4 Observing our Past: Cosmic Archaeology

Our Universe is the simplest possible on two counts: it satisfies the cosmological principle, and it has a flat geometry. The mathematical description of an expanding, homogeneous, and isotropic Universe with a flat geometry is straightforward. We can imagine filling up space with clocks that are all synchronized. At any given snapshot in time the physical conditions (density, temperature) are the same everywhere. But as time goes on, the spatial separation between the clocks will increase. The stretching of space can be described by a time-dependent scale factor, $a(t)$. A separation measured at time t_1 as $r(t_1)$ will appear at time t_2 to have a length $r(t_2) = r(t_1)[a(t_2)/a(t_1)]$.

A natural question to ask is whether our human bodies or even the solar system, are also expanding as the Universe expands. The answer is no, because these systems are held together by forces whose strength far exceeds the cosmic force. The mean density of the Universe today, $\bar{\rho}$, is 29 orders of magnitude smaller than the density of our body. Not only are the electromagnetic forces that keep the atoms in our body together far greater than gravity, but even the gravitational self-force of our body on itself overwhelms the cosmic influence. Only on very large scales does the cosmic gravitational force dominate the scene. This also implies that we cannot observe the cosmic expansion with a local laboratory experiment; in order to notice the expansion we need to observe sources which are spread over the vast scales of millions of light years.

Einstein's equations relate the geometry of space to its matter content. Recent data indicates that our observable section of the Universe is flat (meaning that the sum of the angles in a triangle is 180°). The inferred flatness is a natural consequence of the early period of vast expansion, known as cosmic inflation, during which any initial curvature was flattened. Indeed a small patch of a fixed size (representing our current observable region in the cosmological context) on the surface of a vastly inflated balloon would appear nearly flat. The sum of the angles on a non-expanding triangle placed on this patch would get arbitrarily close to 180 degrees as the balloon inflates.

Einstein's general relativity (GR) equations do not admit a stable steady-state (non-expanding or contracting) solution. A decade after Einstein's invention of

GR, Hubble demonstrated that our Universe is indeed expanding. The space-time of an expanding, homogeneous and isotropic, flat Universe can be described very simply. Because the cosmological principle, we can establish a unique time coordinate throughout space by distributing clocks which are all synchronized throughout the Universe, so that each clock would measure the same time t since the Big Bang. The space-time (4–dimensional) line element ds, commonly defined to vanish for a photon, is described by the Friedmann-Robertson-Walker (FRW) metric,

$$ds^2 = c^2 dt^2 - d\ell^2, \tag{1.1}$$

where c is the speed of light and $d\ell$ is the spatial line-element. The cosmic expansion can be incorporated through a scale factor $a(t)$ which multiples the fixed (x, y, z) coordinates tagging the clocks which are themselves "comoving" with the cosmic expansion. For a flat space,

$$d\ell^2 = a(t)^2(dx^2 + dy^2 + dz^2) = a^2(t)(dR^2 + R^2 d\Omega), \tag{1.2}$$

where $d\Omega = d\theta^2 + \sin^2\theta d\phi^2$ with (R, θ, ϕ) being the spherical coordinates centered on the observer, and $(x, y, z) = R(\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)$

A source located at a separation $r = a(t)R$ from us would move at a velocity $v = dr/dt = \dot{a}R = (\dot{a}/a)r$, where $\dot{a} = da/dt$. Here r is a time-independent tag, denoting the present-day distance of the source. Defining $H = \dot{a}/a$ which is constant in space, we recover the Hubble expansion law $v = Hr$.

Edwin Hubble measured the expansion of the Universe using the Doppler effect. We are all familiar with the same effect for sound waves: when a moving car sounds its horn, the pitch (frequency) we hear is different if the car is approaching us or receding away. Similarly, the wavelength of light depends on the velocity of the source relative to us. As the Universe expands, a light source will move away from us and its Doppler effect will change with time. The Doppler formula for a nearby source of light (with a recession speed much smaller than the speed of light) gives

$$\frac{\Delta\nu}{\nu} \approx -\frac{\Delta v}{c} = -\left(\frac{\dot{a}}{a}\right)\left(\frac{r}{c}\right) = -\frac{(\dot{a}\Delta t)}{a} = -\frac{\Delta a}{a}, \tag{1.3}$$

with the solution, $\nu \propto a^{-1}$. Correspondingly, the wavelength scales as $\lambda = (c/\nu) \propto a$. We could have anticipated this outcome since a wavelength can be used as a measure of distance and should therefore be stretched as the Universe expands. The redshift z is defined through the factor $(1 + z)$ by which the photon wavelength was stretched (or its frequency reduced) between its emission and observation times. If we define $a = 1$ today, then $a = 1/(1 + z)$ at earlier times. Higher redshifts correspond to a higher recession speed of the source relative to us (ultimately approaching the speed of light when the redshift goes to infinity), which in turn implies a larger distance (ultimately approaching our horizon, which is the distance traveled by light since the Big Bang) and an earlier emission time of the source in order for the photons to reach us today.

We see high-redshift sources as they looked at early cosmic times. Observational cosmology is like archaeology – the deeper we look into space the more ancient the clues about our history are (see Figure 1.3). But there is a limit to how far back we can see. In principle, we can image the Universe only as long as it was transparent,

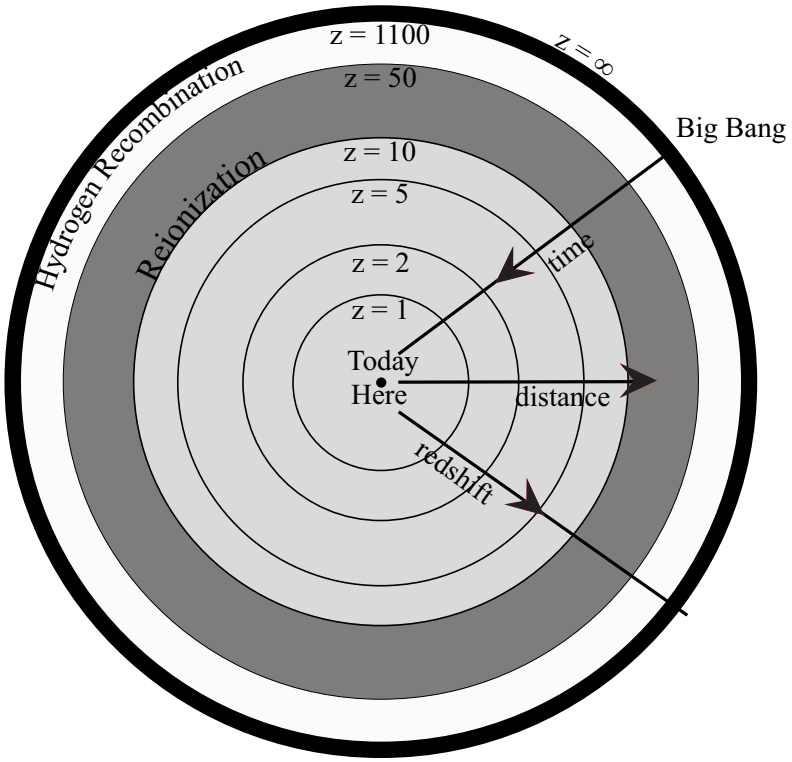

Figure 1.3 Cosmic archaeology of the observable volume of the Universe, in comoving coordinates (which factor out the cosmic expansion). The outermost observable boundary ($z = \infty$) marks the comoving distance that light has traveled since the Big Bang. Future observatories aim to map most of the observable volume of our Universe, and improve dramatically the statistical information we have about the density fluctuations within it. Existing data on the CMB probes mainly a very thin shell at the hydrogen recombination epoch ($z \sim 10^3$, beyond which the Universe is opaque), and current large-scale galaxy surveys map only a small region near us at the center of the diagram. The formation epoch of the first galaxies that culminated with hydrogen reionization at a redshift $z \sim 10$ is shaded grey. Note that the comoving volume out to any of these redshifts scales as the distance cubed.

corresponding to redshifts $z < 10^3$ for photons. The first galaxies are believed to have formed long after that.

The expansion history of the Universe is captured by the scale factor $a(t)$. We can write a simple equation for the evolution of $a(t)$ based on the behavior of a small region of space. For that purpose we need to incorporate the fact that in Einstein's theory of gravity, not only does mass density ρ gravitate but pressure p does too. In a homogeneous and isotropic Universe, the quantity $\rho_{\rm grav} = (\rho + 3p/c^2)$ plays the role of the gravitating mass density ρ of Newtonian gravity.[2] There are several examples to consider. For a radiation fluid,[vi] $p_{\rm rad}/c^2 = \frac{1}{3}\rho_{\rm rad}$, implying that $\rho_{\rm grav} = 2\rho_{\rm rad}$. On the other hand, for a constant vacuum density (the so-called "cosmological constant"), the pressure is negative because by opening up a new volume increment ΔV one gains an energy $\rho c^2 \Delta V$ instead of losing energy, as is the case for normal fluids that expand into more space. In thermodynamics, pressure is derived from the deficit in energy per unit of new volume, which in this case gives $p_{\rm vac}/c^2 = -\rho_{\rm vac}$. This in turn leads to another reversal of signs, $\rho_{\rm grav} = (\rho_{\rm vac} + 3p_{\rm vac}/c^2) = -2\rho_{\rm vac}$, which may be interpreted as repulsive gravity! This surprising result gives rise to the phenomenon of accelerated cosmic expansion, which characterized the early period of cosmic inflation as well as the latest six billions years of cosmic history.

As the Universe expands and the scale factor increases, the matter mass density declines inversely with volume, $\rho_{\rm matter} \propto a^{-3}$, whereas the radiation energy density (which includes the CMB and three species of relativistic neutrinos) decreases as $\rho_{\rm rad}c^2 \propto a^{-4}$, because not only is the density of photons diluted as a^{-3}, but the energy per photon $h\nu = hc/\lambda$ (where h is Planck's constant) declines as a^{-1}. Today $\rho_{\rm matter}$ is larger than $\rho_{\rm rad}$ (assuming massless neutrinos) by a factor of $\sim 3,300$, but at $(1+z) \sim 3,300$ the two were equal, and at even higher redshifts the radiation dominated. Since a stable vacuum does not get diluted with cosmic expansion, the present-day $\rho_{\rm vac}$ remained a constant and dominated over $\rho_{\rm matter}$ and $\rho_{\rm rad}$ only at late times (whereas the unstable "false vacuum" that dominated during inflation has decayed when inflation ended).

1.2.5 Luminosity and Angular-Diameter Distances

When we look at our image reflected off a mirror at a distance of 1 meter, we see the way we looked 6 nano-seconds ago, the time it took light to travel to the mirror and back. If the mirror is spaced 10^{19} cm $=$ 3pc away, we will see the way we looked twenty one years ago. Light propagates at a finite speed, so by observing distant regions, we are able to see how the Universe looked like in the past, a light travel time ago (see Figure 1.3). 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. By using telescopes we can simply see the way

[vi] The momentum of each photon is $\frac{1}{c}$ of its energy. The pressure is defined as the momentum flux along one dimension out of three, and is therefore given by $\frac{1}{3}\rho_{\rm rad}c^2$, where $\rho_{\rm rad}$ is the mass density of the radiation.

distant regions appeared at earlier cosmic times. 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 technological challenges to observers.

We can image the Universe only if it is transparent. Earlier than 400 thousand years after the Big Bang, the cosmic gas was sufficiently hot to be fully ionized (i.e., atoms were broken into free nuclei and electrons), and the Universe was opaque due to scattering by the dense fog of free electrons that filled it. Thus, telescopes cannot be used to image the infant Universe at earlier times (at redshifts $> 10^3$). The earliest possible image of the Universe can be seen in the cosmic microwave background, the thermal radiation left over from the transition to transparency (Figure 1.1).

How faint will the earliest galaxies appear to our telescopes? We can easily express the flux observed from a galaxy of luminosity L at a redshift z. The observed flux (energy per unit time per unit telescope area) is obtained by spreading the energy emitted from the source per unit time, L, over the surface area of a sphere whose radius equals to the effective distance of the source,

$$f = \frac{L}{4\pi d_{\rm L}^2}, \tag{1.4}$$

where $d_{\rm L}$ is defined as the *luminosity distance* in cosmology. For a flat Universe, the comoving distance of a galaxy which emitted its photons at a time $t_{\rm em}$ and is observed at time $t_{\rm obs}$ is obtained by summing over infinitesimal distance elements along the path length of a photon, cdt, each expanded by a factor $(1+z)$ to the present time:

$$r_{\rm em} = \int_{t_{\rm em}}^{t_{\rm obs}} \frac{cdt}{a(t)} = \frac{c}{H_0}\int_0^z \frac{dz'}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}, \tag{1.5}$$

where $a = (1+z)^{-1}$. The *angular diameter distance* $d_{\rm A}$, corresponding to the angular diameter $\theta = D/d_{\rm A}$ occupied by a galaxy of size D, must take into account the fact that we were closer to that galaxy[vii] by a factor $(1+z)$ when the photons started their journey at a redshift z, so it is simply given by $d_{\rm A} = r_{\rm em}/(1+z)$. But to find $d_{\rm L}$ we must take account of additional redshift factors.

If a galaxy has an intrinsic luminosity L, then it would emit an energy $Ldt_{\rm em}$ over a time interval $dt_{\rm em}$. This energy is redshifted by a factor of $(1+z)$ and is observed over a longer time interval $dt_{\rm obs} = dt_{\rm em}(1+z)$ after being spread over a sphere of surface area $4\pi r_{\rm em}^2$. Thus, the observed flux would be

$$f = \frac{Ldt_{\rm em}/(1+z)}{4\pi r_{\rm em}^2 dt_{\rm obs}} = \frac{L}{4\pi r_{\rm em}^2(1+z)^2}, \tag{1.6}$$

implying that[viii]

$$d_{\rm L} = r_{\rm em}(1+z) = d_{\rm A}(1+z)^2. \tag{1.7}$$

[vii] In a flat Universe, photons travel along straight lines. The angle at which a photon is seen is not modified by the cosmic expansion, since the Universe expands at the same rate both parallel and perpendicular to the line of sight.

[viii] A simple analytic fitting formula for $d_L(z)$ was derived by Pen, U.-L. *Astrophys. J. Suppl.* **120**, 49 (1999); http://arxiv.org/pdf/astro-ph/9904172v1 .

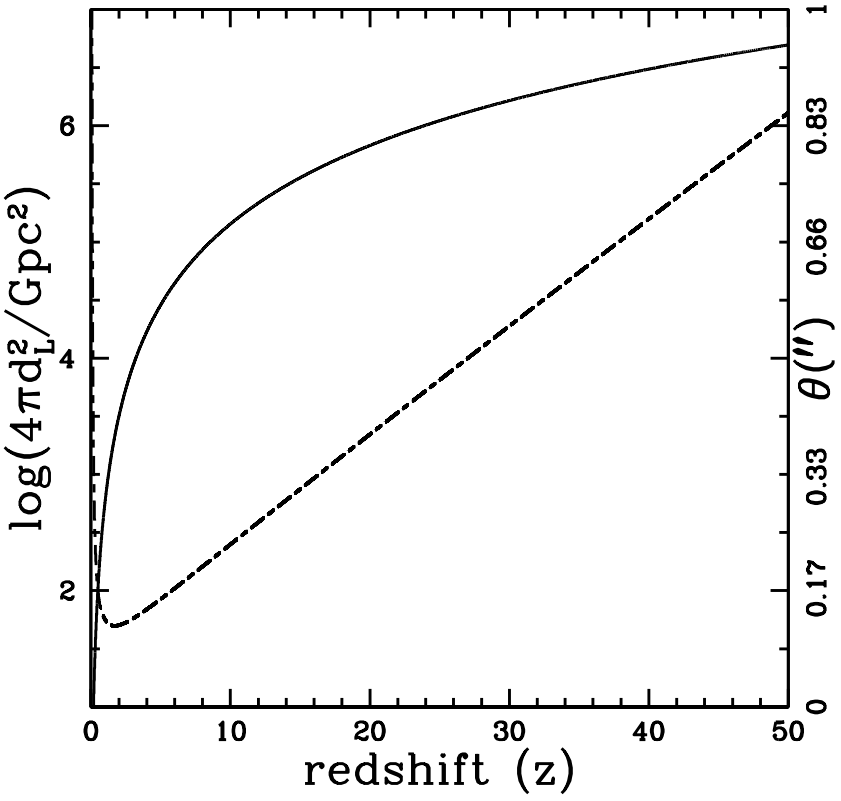

Figure 1.4 The solid line (corresponding to the label on the left-hand side) shows Log_{10} of the conversion factor between the luminosity of a source and its observed flux, $4\pi d_{\mathrm{L}}^2$ (in Gpc^2), as a function of redshift, z. The dashed-dotted line (labeled on the right) gives the angle θ (in arcseconds) occupied by a galaxy of a 1 kpc diameter as a function of redshift.

The area dilution factor $4\pi d_{\rm L}^2$ is plotted as a function of redshift in the bottom panel of Figure 1.4. If the observed flux is only measured over a narrow band of frequencies, one needs to take account of the additional conversion factor of $(1+z) = (d\nu_{\rm em}/d\nu_{\rm obs})$ between the emitted frequency interval $d\nu_{\rm em}$ and its observed value $d\nu_{\rm obs}$. This yields the relation $(df/d\nu_{\rm obs}) = (1+z) \times (dL/d\nu_{\rm em})/(4\pi d_{\rm L}^2)$. Figure **??** compares the predicted flux per unit frequency[ix] from a galaxy at a redshift $z_s = 10$ for a Salpeter IMF and for massive ($> 100 M_\odot$) Population III stars, in units of nJy per $10^6 M_\odot$ in stars (where 1 nJy$= 10^{-32}$ erg cm^{-2} s^{-1} Hz^{-1}). The observed flux is an order of magnitude larger in the Population III case. The strong UV emission by massive stars is likely to produce bright recombination lines, such as Lyman-α and He II 1640 Å, from the interstellar medium surrounding these stars.

Theoretically, the expected number of early galaxies of different fluxes per unit area on the sky can be calculated by dressing up the dark matter halos in Figure 3.4 with stars of some prescribed mass distribution and formation history, then finding the corresponding abundance of galaxies of different luminosities as a function of redshift.[3] There are many uncertain parameters in this approach (such as $f_\star$, $f_{\rm esc}$, the stellar mass function, the star formation time, the metallicity, and feedback), so one is tempted to calibrate these parameters by observing the sky.[4]

1.3 MILESTONES IN COSMIC EVOLUTION

The gravitating mass, $M_{\rm grav} = \rho_{\rm grav} V$, enclosed by a spherical shell of radius $a(t)$ and volume $V = \frac{4\pi}{3} a^3$, induces an acceleration

$$\frac{d^2 a}{dt^2} = -\frac{G M_{\rm grav}}{a^2}. \tag{1.8}$$

Since $\rho_{\rm grav} = \rho + 3p/c^2$, we need to know how pressure evolves with the expansion factor $a(t)$. This is obtained from the thermodynamic relation mentioned above between the change in the internal energy $d(\rho c^2 V)$ and the pdV work done by the pressure, $d(\rho c^2 V) = -pdV$. This relation implies $-3pa\dot{a}/c^2 = a^2\dot{\rho} + 3\rho a\dot{a}$, where a dot denotes a time derivative. Multiplying equation (1.8) by $\dot{a}$ and making use of this relation yields our familiar result

$$E = \frac{1}{2}\dot{a}^2 - \frac{GM}{a}, \tag{1.9}$$

where E is a constant of integration and $M \equiv \rho V$. As discussed before, the spherical shell will expand forever (being gravitationally unbound) if $E \geq 0$, but will eventually collapse (being gravitationally bound) if $E < 0$. Making use of the Hubble parameter, $H = \dot{a}/a$, equation (1.9) can be re-written as

$$\frac{E}{\frac{1}{2}\dot{a}^2} = 1 - \Omega, \tag{1.10}$$

[ix] The observed flux per unit frequency can be translated to an equivalent AB magnitude using the relation, AB $\equiv -2.5 \log_{10}[(df/d\nu_{\rm obs})/{\rm erg\ s^{-1}\ cm^{-2}\ Hz^{-1}}] - 48.6$.

where $\Omega = \rho/\rho_c$, with

$$\rho_c = \frac{3H^2}{8\pi G} = 9.2 \times 10^{-30} \frac{\mathrm{g}}{\mathrm{cm}^3} \left(\frac{H}{70 \ \mathrm{km \ s^{-1} Mpc^{-1}}} \right)^2 . \tag{1.11}$$

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 of protons and neutrons, or "baryons"), vacuum density (cosmological constant), and radiation, respectively, a flat universe satisfies

$$\frac{H(t)}{H_0} = \left[\frac{\Omega_m}{a^3} + \Omega_\Lambda + \frac{\Omega_r}{a^4} \right]^{1/2} , \tag{1.12}$$

where we define H_0 and $\Omega_0 = (\Omega_m + \Omega_\Lambda + \Omega_r) = 1$ to be the present-day values of H and Ω, respectively.

In the particularly simple case of a flat Universe, we find that if matter dominates then $a \propto t^{2/3}$, if radiation dominates then $a \propto t^{1/2}$, and if the vacuum density dominates then $a \propto \exp\{H_{\rm vac} t\}$ with $H_{\rm vac} = (8\pi G \rho_{\rm vac}/3)^{1/2}$ being a constant. In the beginning, after inflation ended, the mass density of our Universe ρ was at first dominated by radiation at redshifts $z > 3,300$, then it became dominated by matter at $0.3 < z < 3,300$, and finally was dominated by the vacuum at $z < 0.3$. The vacuum started to dominate $\rho_{\rm grav}$ already at $z < 0.7$ or six billion years ago. Figure 1.6 illustrates the mass budget in the present-day Universe and during the epoch when the first galaxies had formed.

The above results for $a(t)$ have two interesting implications. First, we can figure out the relationship between the time since the Big Bang and redshift since $a = (1+z)^{-1}$. For example, during the matter-dominated era ($1 < z < 10^3$, with the low-z end set by the condition $[1+z] \gg [\Omega_\Lambda/\Omega_m]^{1/3}$),

$$t \approx \frac{2}{3H_0 {\Omega_m}^{1/2} (1+z)^{3/2}} = \frac{0.95 \times 10^9 \ \mathrm{years}}{[(1+z)/7]^{3/2}}. \tag{1.13}$$

In this same regime, where $\Omega_m \approx 1$, $H \approx 2/(3t)$ and $a = (1+z)^{-1} \approx (3H_0\sqrt{\Omega_m}/2)^{2/3} t^{2/3}$.

Second, we note the remarkable exponential expansion for a vacuum dominated phase. This accelerated expansion serves an important purpose in explaining a few puzzling features of our Universe. We already noticed that our Universe was prepared in a very special initial state: nearly isotropic and homogeneous, with Ω close to unity and a flat geometry. In fact, it took the CMB photons nearly the entire age of the Universe to travel towards us. Therefore, it should take them twice as long to bridge across their points of origin on opposite sides of the sky. *How is it possible then that the conditions of the Universe (as reflected in the nearly uniform CMB temperature) were prepared to be the same in regions that were never in causal contact before?* Such a degree of organization is highly unlikely to occur at random. If we receive our clothes ironed out and folded neatly, we know that there must have a been a process that caused it. Cosmologists have identified an analogous "ironing process" in the form of *cosmic inflation*. This process is associated with an early period during which the Universe was dominated temporarily by the mass density of an elevated vacuum state, and experienced exponential expansion

by at least ~ 60 e-folds. This vast expansion "ironed out" any initial curvature of our environment, and generated a flat geometry and nearly uniform conditions across a region far greater than our current horizon. After the elevated vacuum state decayed, the Universe became dominated by radiation.

The early epoch of inflation is important not just in producing the global properties of the Universe but also in generating the inhomogeneities that seeded the formation of galaxies within it.[5] The vacuum energy density that had driven inflation encountered quantum mechanical fluctuations. After the perturbations were stretched beyond the horizon of the infant Universe (which today would have occupied the size no bigger than a human hand), they materialized as perturbations in the mass density of radiation and matter. The last perturbations to leave the horizon during inflation eventually entered back after inflation ended (when the scale factor grew more slowly than ct). It is tantalizing to contemplate the notion that galaxies, which represent massive classical objects with $\sim 10^{67}$ atoms in today's Universe, might have originated from sub-atomic quantum-mechanical fluctuations at early times.

After inflation, an unknown process, called "baryo-genesis" or "lepto-genesis", generated an excess of particles (baryons and leptons) over anti-particles.[x] As the Universe cooled to a temperature of hundreds of MeV (with $1\mathrm{MeV}/k_B = 1.1604 \times 10^{10}\mathrm{K}$), protons and neutrons condensed out of the primordial quark-gluon plasma through the so-called *QCD phase transition*. At about one second after the Big Bang, the temperature declined to ~ 1 MeV, and the weakly interacting neutrinos decoupled. Shortly afterwards the abundance of neutrons relative to protons froze and electrons and positrons annihilated. In the next few minutes, nuclear fusion reactions produced light elements more massive than hydrogen, such as deuterium, helium, and lithium, in abundances that match those observed today in regions where gas has not been processed subsequently through stellar interiors. Although the transition to matter domination occurred at a redshift $z \sim 3,300$ the Universe remained hot enough for the gas to be ionized, and electron-photon scattering effectively coupled ordinary matter and radiation. At $z \sim 1,100$ the temperature dipped below $\sim 3,000$K, and free electrons recombined with protons to form neutral hydrogen atoms. As soon as the dense fog of free electrons was depleted, the Universe became transparent to the relic radiation, which is observed at present as the CMB. These milestones of the thermal history are depicted in Figure 1.5.

The Big Bang is the only known event in our past history where particles interacted with center-of-mass energies approaching the so-called "Planck scale"[xi] $[(hc^5/G)^{1/2} \sim 10^{19}\ \mathrm{GeV}]$, at which quantum mechanics and gravity are expected to be unified. Unfortunately, the exponential expansion of the Universe during inflation erased memory of earlier cosmic epochs, such as the Planck time.

---

[x] Anti-particles are identical to particles but with opposite electric charge. Today, the ordinary matter in the Universe is observed to consist almost entirely of particles. The origin of the asymmetry in the cosmic abundance of matter over anti-matter is stil an unresolved puzzle.

[xi] The Planck energy scale is obtained by equating the quantum-mechanical wavelength of a relativistic particle with energy E, namely hc/E, to its "black hole" radius $\sim GE/c^4$, and solving for E.

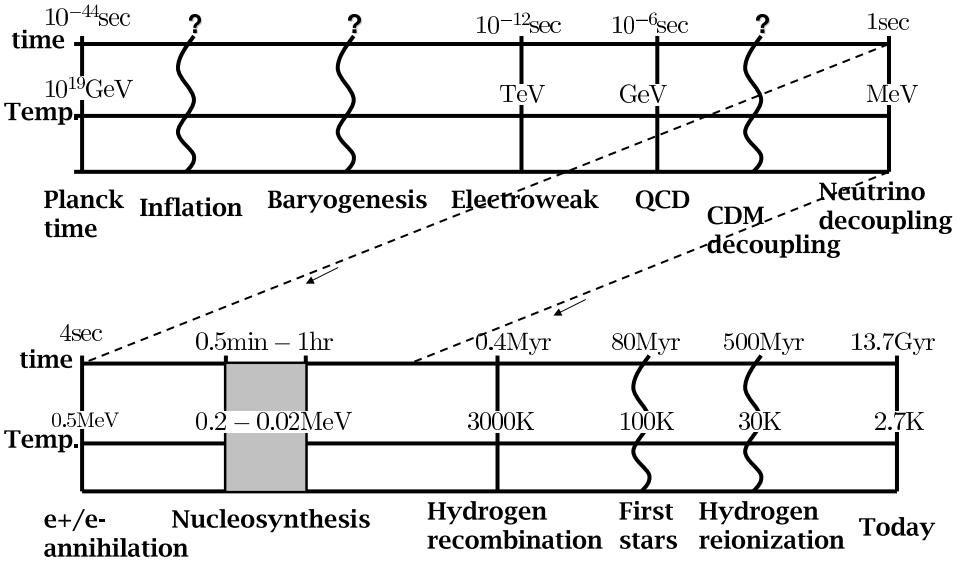

Figure 1.5 Following inflation, the Universe went through several other milestones which left a detectable record. These include baryogenesis (which resulted in the observed asymmetry between matter and anti-matter), the electroweak phase transition (during which the symmetry between electromagnetic and weak interactions was broken), the QCD phase transition (during which protons and neutrons nucleated out of a soup of quarks and gluons), the dark matter decoupling epoch (during which the dark matter decoupled thermally from the cosmic plasma), neutrino decoupling, electron-positron annihilation, light-element nucleosynthesis (during which helium, deuterium and lithium were synthesized), and hydrogen recombination. The cosmic time and CMB temperature of the various milestones are marked. Wavy lines and question marks indicate milestones with uncertain properties. The signatures that the same milestones left in the Universe are used to constrain its parameters.

1.4 MOST MATTER IS DARK

Surprisingly, most of the matter in the Universe is not the same ordinary matter that we are made of (see Figure 1.6). If it were ordinary matter (which also makes stars and diffuse gas), it would have interacted with light, thereby revealing its existence to observations through telescopes. Instead, observations of many different astrophysical environments require the existence of some mysterious dark component of matter which only reveals itself through its gravitational influence and leaves no other clue about its nature. Cosmologists are like a detective who finds evidence for some unknown criminal in a crime scene and is anxious to find his/her identity. The evidence for dark matter is clear and indisputable, assuming that the laws of gravity are not modified (although a small minority of scientists are exploring this alternative).

Without dark matter we would have never existed by now. This is because ordinary matter is coupled to the CMB radiation that filled up the Universe early on. The diffusion of photons on small scales smoothed out perturbations in this primordial radiation fluid. The smoothing length was stretched to a scale as large as hundreds of millions of light years in the present-day Universe. This is a huge scale by local standards, since galaxies – like the Milky Way – were assembled out of matter in regions a hundred times smaller than that. Because ordinary matter was coupled strongly to the radiation in the early dense phase of the Universe, it also was smoothed on small scales. If there was nothing else in addition to the radiation and ordinary matter, then this smoothing process would have had a devastating effect on the prospects for life in our Universe. Galaxies like the Milky Way would have never formed by the present time since there would have been no density perturbations on the relevant small scales to seed their formation. The existence of dark matter not coupled to the radiation came to the rescue by keeping memory of the initial seeds of density perturbations on small scales. In our neighborhood, these seed perturbations led eventually to the formation of the Milky Way galaxy inside of which the Sun was made as one out of tens of billions of stars, and the Earth was born out of the debris left over from the formation process of the Sun. This sequence of events would have never occurred without the dark matter.

We do not know what the dark matter is made of, but from the good match obtained between observations of large-scale structure and the equations describing a pressureless fluid (see equations 2.3-2.4), we infer that it is likely made of particles with small random velocities. It is therefore called "cold dark matter" (CDM). The popular view is that CDM is composed of particles which possess weak interactions with ordinary matter, similarly to the elusive neutrinos we know to exist. The abundance of such particles would naturally "freeze-out" at a temperature $T > 1$MeV, when the Hubble expansion rate is comparable to the annihilation rate of the CDM particles. Interestingly, such a decoupling temperature naturally leads through a Boltzmann suppression factor $\sim \exp\{-mc^2/k_BT\}$ to Ω_m of order unity for particle masses of $mc^2 > 100$ GeV with a weak interaction cross-section, as expected for the lightest (and hence stable) supersymmetric particle in simple extensions of the standard model of particle physics. The hope is that CDM particles, owing to their weak but non-vanishing coupling to ordinary matter, will nevertheless be

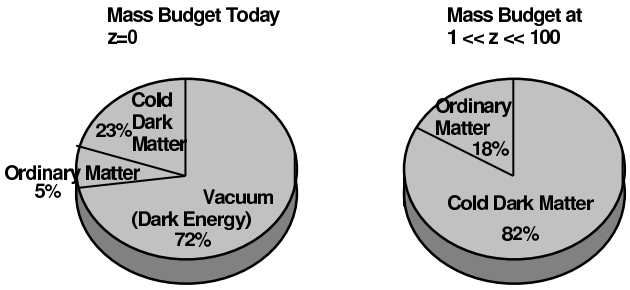

Figure 1.6 Mass budgets of different components in the present day Universe and in the infant Universe when the first galaxies formed (redshifts $z = 10$–50). The CMB radiation (not shown) makes up a fraction $\sim 0.03\%$ of the budget today, but was dominant at redshifts $z > 3,300$. The cosmological constant (vacuum) contribution was negligible at high redshifts ($z \gg 1$).

produced in small quantities through collisions of energetic particles in future laboratory experiments such as the Large Hadron Collider (LHC).[6] Other experiments are attempting to detect directly the astrophysical CDM particles in the Milky Way halo. A positive result from any of these experiments will be equivalent to our detective friend being successful in finding a DNA sample of the previously unidentified criminal.

The most popular candidate for the cold dark matter (CDM) particle is a Weakly Interacting Massive Particle (WIMP). The lightest supersymmetric particle (LSP) could be a WIMP. The CDM particle mass depends on free parameters in the particle physics model; the LSP hypothesis will be tested at the Large Hadron Collider or in direct detection experiments. The properties of the CDM particles affect their response to the primordial inhomogeneities on small scales. The particle cross-section for scattering off standard model particles sets the epoch of their thermal decoupling from the cosmic plasma.

The dark ingredients of the Universe can only be probed indirectly through a variety of luminous tracers. The distribution and nature of the dark matter are constrained by detailed X-ray and optical observations of galaxies and galaxy clusters. The evolution of the dark energy with cosmic time will be constrained over the coming decade by surveys of Type Ia supernovae, as well as surveys of X-ray clusters, up to a redshift of two.

According to the standard cosmological model, the CDM behaves as a collection of collisionless particles that started out at the epoch of matter domination

with negligible thermal velocities, and later 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. 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 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 (when protons and electrons combined to make neutral hydrogen), 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 reionized.

The initial conditions of the Universe can be summarized on a single sheet of paper. The small number of parameters that provide an accurate statistical description of these initial conditions are summarized in Table 1.1. However, thousands of books in libraries throughout the world cannot summarize the complexities of galaxies, stars, planets, life, and intelligent life, in the present-day Universe. If we feed the simple initial cosmic conditions into a gigantic computer simulation incorporating the known laws of physics, we should be able to reproduce all the complexity that emerged out of the simple early universe. Hence, all the information associated with this later complexity was encapsulated in those simple initial conditions. Below we follow the process through which late time complexity appeared and established an irreversible arrow to the flow of cosmic time.[xii]

The basic question that cosmology attempts to answer is: **What is the composition of the Universe and what initial conditions generated the observed structures in it?** In detail, we would like to know:
(a) Did inflation occur and when? If so, what drove it and how did it end?
(b) What is the nature of of the dark energy and how does it change over time and

[xii] In previous decades, astronomers used to associate the simplicity of the early Universe with the fact that the data about it was scarce. Although this was true at the infancy of observational cosmology, it is not true any more. With much richer data in our hands, the initial simplicity is now interpreted as an outcome of inflation.

Table 1.1 Standard set of cosmological parameters (defined and adopted throughout the book). Based on Komatsu,E., et al. *Astrophys. J. Suppl.* **180**, 330 (2009).

Ω_Λ	Ω_m	Ω_b	h	n_s	σ_8
0.72	0.28	0.05	0.7	1	0.82

space?
(c) What is the nature of the dark matter and how did it regulate the evolution of structure in the Universe?

The first galaxies were shaped, more than any other class of astrophysical objects, by the pristine initial conditions and basic constituents of the Universe. Studying the formation process of the first galaxies could reveal unique evidence for new physics that was so far veiled in older galaxies by complex astrophysical processes.

Chapter Two

From Recombination to the First Galaxies

After cosmological recombination, the Universe entered the "dark ages" during which the relic CMB light from the Big Bang gradually faded away. During this "pregnancy" period which lasted hundreds of millions of years, the seeds of small density fluctuations planted by inflation in the matter distribution grew up until they eventually collapsed to make the first galaxies.[7]

2.1 GROWTH OF LINEAR PERTURBATIONS

As discussed earlier, small perturbations in density grow due to the unstable nature of gravity. Overdense regions behave as if they reside in a closed Universe. Their evolution ends in a "big crunch", which results in the formation of gravitationally bound objects like the Milky Way galaxy.

Equation (1.10) explains the formation of galaxies out of seed density fluctuations in the early Universe, at a time when the mean matter density was very close to the critical value and $\Omega_m \approx 1$. Given that the mean cosmic density was close to the threshold for collapse, a spherical region which was only slightly denser than the mean behaved as if it was part of an $\Omega > 1$ universe, and therefore eventually collapsed to make a bound object, like a galaxy. The material from which objects are made originated in the underdense regions (voids) that separate these objects (and which behaved as part of an $\Omega < 1$ Universe), as illustrated in Figure 1.2.

Observations of the CMB show that at the time of hydrogen recombination the Universe was extremely uniform, with spatial fluctuations in the energy density and gravitational potential of roughly one part in 10^5. These small fluctuations grew over time during the matter dominated era as a result of gravitational instability, and eventually led to the formation of galaxies and larger-scale structures, as observed today.

In describing the gravitational growth of perturbations in the matter-dominated era ($z \ll 3,300$), we may consider small perturbations of a fractional amplitude $|\delta| \ll 1$ on top of the uniform background density $\bar{\rho}$ of cold dark matter. The three fundamental equations describing conservation of mass and momentum along with the gravitational potential can then be expanded to leading order in the perturbation amplitude. We distinguish between physical and comoving coordinates (the latter expanding with the background Universe). Using vector notation, the fixed coordinate $\mathbf{r}$ corresponds to a comoving position $\mathbf{x} = \mathbf{r}/a$. 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 fractional density perturbations

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{r})}{\bar{\rho}} - 1 \ , \tag{2.1}$$

where the mean fluid mass density is $\bar{\rho}$, with a corresponding peculiar velocity which describes the deviation from the Hubble flow $\mathbf{u} \equiv \mathbf{v} - H\mathbf{r}$. The fluid is then described by the continuity and Euler equations in comoving coordinates:

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot [(1+\delta)\mathbf{u}] = 0 \tag{2.2}$$

$$\frac{\partial \mathbf{u}}{\partial t} + H\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{a}\nabla\phi \ . \tag{2.3}$$

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

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \ . \tag{2.4}$$

This fluid description is valid for describing the evolution of collisionless cold dark matter particles until different particle streams cross. The crossing typically occurs only after perturbations have grown to become non-linear with $|\delta| > 1$, and at that point the individual particle trajectories must in general be followed.

The combination of the above equations yields to leading order in δ,

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta \ . \tag{2.5}$$

This linear equation has in general two independent solutions, only one of which grows in 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 amplitude in proportion to a growth factor $D(t)$. The growth factor in a flat (matter-dominated) Universe at $z < 10^3$ is given by[i]

$$D(t) \propto \frac{\left(\Omega_\Lambda a^3 + \Omega_m\right)^{1/2}}{a^{3/2}} \int_0^a \frac{a'^{3/2}\, da'}{\left(\Omega_\Lambda a'^3 + \Omega_m\right)^{3/2}} \ . \tag{2.6}$$

In the matter-dominated regime of the redshift range $1 < z < 10^3$, the growth factor is simply proportional to the scale factor $a(t)$. Interestingly, the gravitational potential $\phi \propto \delta/a$ does not grow in comoving coordinates. This implies that the potential depth fluctuations remain frozen in amplitude as fossil relics from the inflationary epoch during which they were generated. Nonlinear collapse only changes the potential depth by a factor of order unity, but even inside collapsed objects its rough magnitude remains as testimony to the inflationary conditions. This explains why the characteristic potential depth of collapsed objects such as galaxy clusters ($\phi/c^2 \sim 10^{-5}$) is of the same order as the potential fluctuations probed by the fractional variations in the CMB temperature across the sky. At low redshifts $z < 1$ and in the future, the cosmological constant dominates ($\Omega_m \ll \Omega_\Lambda$) and

[i] An analytic expression for the growth factor in terms of special functions was derived by Eisenstein, D. (1997), http://arxiv.org/pdf/astro-ph/9709054v2 .

the density fluctuations freeze in amplitude ($D(t) \rightarrow$constant) as their growth is suppressed by the accelerated expansion of space.

It is also useful to consider the velocity field $\mathbf{u}$. To analyze its evolution, it is convenient to express the density field as a sum over a complete set of periodic "Fourier modes," each having a sinusoidal (wave-like) dependence on space with a comoving wavelength $\lambda = 2\pi/k$ and wavenumber k. Mathematically, we write[ii]

$$\delta_{\mathbf{k}} = \int d^3x\, \delta(x) e^{i\mathbf{k}\cdot\mathbf{x}} \tag{2.7}$$

$$\delta(\mathbf{x}) = \int \frac{d^3k}{(2\pi)^3} \delta_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{2.8}$$

with $\mathbf{x}$ being the comoving spatial coordinate. The characteristic amplitude of each $\mathbf{k}$-mode defines the typical value of δ on the spatial scale λ. It is easy to show that equation (2.5) applies to each Fourier mode individually, so the factor $D(t)$ also describes their growth (in the linear regime), and the evolution of the density field in Fourier space is easy to follow. In particular, note that different spatial scales evolve *independently* in the linear regime.

To linear order, the continuity equation (2.2) becomes $\nabla \cdot \mathbf{u} = -a(d\delta/dt)$, or in Fourier space

$$-i\mathbf{k} \cdot \mathbf{u}_{\mathbf{k}} = -\frac{a}{D}\frac{dD}{dt}\delta_{\mathbf{k}}, \tag{2.9}$$

where we have assumed that $\delta_{\mathbf{k}}$ is a pure growing mode. This has the solution

$$\mathbf{u}_{\mathbf{k}} = -i\frac{aHf(\Omega)}{k}\delta_{\mathbf{k}}\hat{\mathbf{k}}, \tag{2.10}$$

where $f(\Omega) = (a/D)(dD/da) \approx \Omega_m^{0.6}$ to a very good approximation (note that it is almost independent of Ω_Λ). Interestingly, peculiar velocity perturbations grow proportionally to density fluctuations, and their growing modes are parallel to the wavevector. Note also that $\mathbf{u}_k b \propto \delta_{\mathbf{k}}/k$, which implies that peculiar velocities are sourced by gravitational fluctuations on *larger* scales than those of the density field.

2.1.1 The Power Spectrum of Density Fluctuations

The initial perturbation amplitude varies with spatial scale; typically, large-scale regions have a smaller perturbation amplitude than small-scale regions. The statistical properties of the perturbations as a function of spatial scale can be best captured by its Fourier transform in comoving wavenumbers. This approach has the convenient property that the spatial scales are *fixed* in time, rather than evolve as the perturbation expands or collapses.

Because we cannot observe particular regions mature and grow over time, we are typically concerned not with the amplitude of particular density perturbations or modes but with the properties of their statistical ensemble. There are two complementary statistical measures that are used most often. The first is the *correlation function*,

$$\xi(\mathbf{x}) = \langle \delta(\mathbf{x})\delta(0) \rangle\,, \tag{2.11}$$

[ii] Note that cosmologists typically absorb the volume factors in the Fourier transform into $\delta_{\mathbf{k}}$, which has units of volume.

where the angular brackets represent averaging over the entire statistical ensemble of points separated by a comoving distance $\mathbf{x}$, and where we made use of the translational invariance of statistical averages in centering our coordinate system on the second point. The correlation function expresses the degree to which a particular overdensity is more likely to be surrounded by other overdense regions. Note that for an isotropic distribution of perturbations, ξ is a function only of the magnitude of the spatial separation, $x = |\mathbf{x}|$.

The second is the *power spectrum*,

$$P(\mathbf{k}) = \langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \rangle = (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') P(\mathbf{k}), \tag{2.12}$$

which has units of volume. Again, it is a function only of $k = |\mathbf{k}|$ for an isotropic universe.

In fact, the correlation function and power spectrum are intimately related. If we write the former using the Fourier transform of $\delta(\mathbf{x})$, we obtain

$$\xi(\mathbf{x}) = \left\langle \int \frac{d^3k}{(2\pi)^3} \delta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \int \frac{d^3k'}{(2\pi)^3} \delta^*_{\mathbf{k}'} \right\rangle \tag{2.13}$$

$$= \int \frac{d^3k}{(2\pi)^3} \int \frac{d^3k'}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} \langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \rangle \tag{2.14}$$

$$= \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} P(\mathbf{k}), \tag{2.15}$$

where in the first line we have used the fact that $\delta(0)$ is real. Thus $\xi(r)$ and $P(k)$ are simply Fourier transforms of each other. Theoretical calculations are generally simplest using the Fourier representation and power spectrum, but the two approaches have different error properties so both are used regularly in the literature.

Inflation generates perturbations in which different $\mathbf{k}$-modes are statistically independent, and each has a random phase constant in its sinusoid. The statistical properties of these fluctuations are perfectly described by the power spectrum. Moreover, in the standard cosmological model, inflation produces a very simple primordial power-law spectrum $P(k) \propto k^{n_s}$ with $n_s \approx 1$. This spectrum admits the special property that gravitational potential fluctuations of all wavelengths have the same amplitude at the time when they enter the horizon (namely, when their wavelength matches the distance traveled by light during the age of the Universe), and so this spectrum is called "scale-invariant." This spectrum has the aesthetic appeal that perturbations can always be small on the horizon scale. A different power-law spectrum would either lead to an overdensity of order unity across the horizon, resulting in black hole formation, either in the Universe's future or past. Quantum fluctuations during cosmic inflation naturally results in a nearly scale-invariant spectrum because of the near constancy of the Hubble parameter for a nearly steady vacuum density.

However, the power spectrum becomes more complex as perturbations grow at later times in a CDM universe. In particular, the modified final power spectrum is characterized by a turnover at a scale of order the horizon cH^{-1} at matter-radiation equality, and a small-scale asymptotic shape of $P(k) \propto k^{n_s-4}$. The turnover results from the fact that density perturbations experience almost no growth during

the radiation dominated era, because the Jeans length then ($\sim ct/\sqrt{3}$; see the next chapter) is comparable to the scale of the horizon inside of which growth is enabled by causality. Therefore, modes on a spatial scale that entered the horizon during the early radiation-dominated era show a smaller amplitude relative to the power-law extrapolation of long wavelength modes that entered the horizon during the matter-dominated era. For a scale-invariant index $n_s \approx 1$, the small-scale fluctuations have the same amplitude at horizon crossing, and with nearly no growth they have the same amplitude on all sub-horizon mass scales at matter-radiation equality. The associated constancy of the fluctuation amplitude on small mass scales (in real space), $\delta^2 \propto P(k)k^3 \sim const$, implies a small-scale asymptotic slope for $P(k)$ of ≈ -3 or $(n_s - 4)$. The resulting power-spectrum after matter-radiation equality is crudely described by the fitting function,[8]

$$P(k) \propto k^{n_s}/(1 + \alpha_p k + \beta_p k^2)^2, \tag{2.16}$$

with $\alpha_p = 8(\Omega_m h^2)^{-1}$ Mpc and $\beta_p = 4.7(\Omega_m h^2)^{-2}$ Mpc^2, and refinements that depend on the baryon mass fraction and neutrino properties (mass and number of flavors).[9]

Species that decouple at a particular time from the cosmic plasma (like the dark matter or the baryons) would show fossil evidence for acoustic oscillations in their power spectrum of inhomogeneities due to sound waves in the radiation fluid to which they were coupled at early times. This phenomenon can be understood as follows. Imagine a localized point-like perturbation from inflation at $t = 0$. The small perturbation in density or pressure will send out a sound wave that will reach the sound horizon $c_s t$ at any later time t. The perturbation will therefore correlate with its surroundings up to the sound horizon and all k-modes with wavelengths equal to this scale or its harmonics will be correlated. This results in a series of peaks in the power-spectrum corresponding to these harmonics. The peaks in the power spectrum of the baryons after recombination induce corresponding peaks in the dark matter sector at later times. These peaks are on spatial scales far greater than the intrinsic peaks associated with the much earlier decoupling epoch of the dark matter (which for weakly-interacting particles correspond to mass scales of planets or smaller). The mass scales of the perturbations that grow to become the first collapsed objects at $z < 100$ cross the horizon in the radiation dominated era after the dark matter had already decoupled from the cosmic plasma.

Although this shape is well determined by linear perturbation theory in an expanding universe, the overall *amplitude* of the power spectrum is not specified by current models of inflation, and is usually set by comparing to the observed CMB temperature fluctuations or to measures of large-scale structure based on surveys of galaxies, clusters of galaxies, or the intergalactic gas. Computer codes that compute the detailed shape of the power-spectrum are publicly available at **http://camb.info/** and **http://www.cmbfast.org**

The most popular large-scale structure normalization is through the observed mass fluctuation amplitude (at the present day) on $8h^{-1}$ Mpc, roughly the scale of galaxy clusters. To relate this quantity to the power spectrum, we must consider the statistical distribution of the smoothed density field. We define a window (or filter) function $W(\mathbf{r})$ normalized so that $\int d^3r\, W(\mathbf{r}) = 1$, with the smoothed den-

sity perturbation field being $\int d^3r\delta(\mathbf{x})W(\mathbf{r})$. The simplest observed quantity is to measure masses (relative to the mean) inside spheres of radius R; in this case we use a "spherical top-hat" window (similar to a three-dimensional cookie cutter), in which $W =$ constant inside a sphere of radius R and $W = 0$ outside.

The normalization of the present power spectrum at $z = 0$ is then specified by the variance of this density field when smoothed on the particular scale of $8h^{-1}\mathrm{Mpc}$, $\sigma_8 \equiv \sigma(R = 8h^{-1}\mathrm{Mpc})$. For the top-hat filter, the smoothed perturbation field is denoted by δ_R or δ_M, where the enclosed 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. We then write the variance $\left\langle \delta_M^2 \right\rangle$ (relative to the mean) as[iii]

$$\sigma^2(M) = \left\langle \frac{1}{V}\int d^3x\delta(\mathbf{x})W(\mathbf{x})\,\frac{1}{V}\int d^3x'\delta(\mathbf{x}')W(\mathbf{x}')\right\rangle \tag{2.17}$$

$$= \frac{1}{V^2}\int d^3x\, d^3x'\, W(\mathbf{x})W(\mathbf{x}')\xi(|\mathbf{x}-\mathbf{x}'|) \tag{2.18}$$

$$= \int \frac{d^3k}{(2\pi)^3}P(k)\frac{|W_\mathbf{k}|^2}{V^2}, \tag{2.19}$$

where $W_\mathbf{k}$ is the Fourier transform of the window function. For the usual choice of a spherical top hat, this is

$$\sigma^2(M) \equiv \sigma^2(R) = \int_0^\infty \frac{dk}{k}\,\Delta^2(k)\left[\frac{3j_1(kR)}{kR}\right]^2, \tag{2.20}$$

where $j_1(x) = (\sin x - x\cos x)/x^2$ and $\Delta^2(k) = k^3P(k)/2\pi^2$ is the so-called dimensionless power spectrum. Δ^2 expresses the contribution, per log wavenumber, of the power spectrum to the net variance.

While the normalization of the power spectrum only requires σ_8, we will see in the next chapter that the function $\sigma(M)$ plays a major role in fixing the abundance of collapsed objects. We therefore show it in Figure 2.1 as a function of mass and redshift for the standard cosmological model. Note that $\sigma^2 \propto \delta^2 \propto D(t)^2$, so the time dependence is trivial (at least in linear theory).

For modes with random phases, the probability of different regions with the same comoving size M to have a perturbation amplitude between δ and $\delta+d\delta$ is Gaussian with a zero mean and a variance $\sigma^2(M)$,

$$P(\delta)d\delta = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\delta^2/2\sigma^2}\,d\delta. \tag{2.21}$$

These so-called Gaussian perturbations are a key prediction of inflation; they have the convenient property that the statistical distribution of densities is described entirely by the power spectrum (through σ^2). A very small amount of primordial non-gaussianity can be accommodated; the nonlinear phase of gravitational collapse generates more.

[iii] Note that σ^2 can equally well be considered a function of spatial scale R.

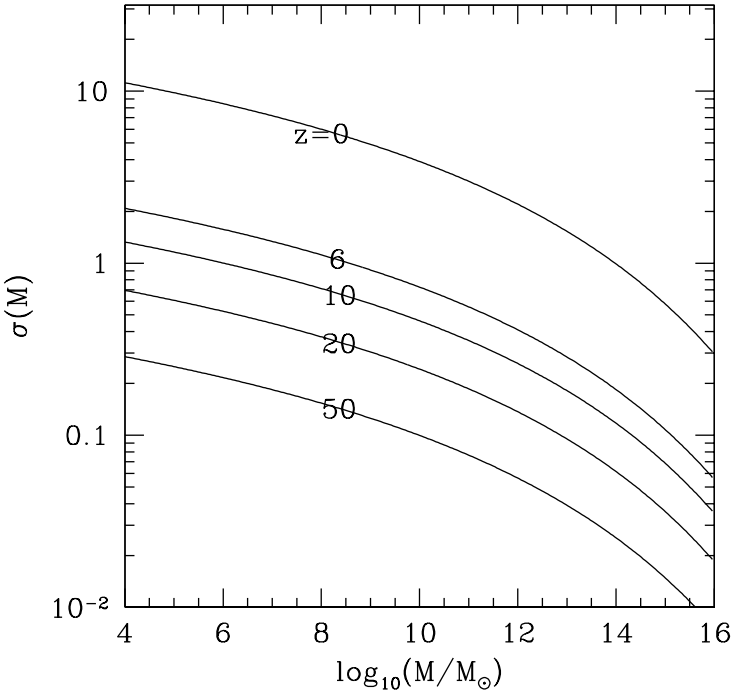

Figure 2.1 The root-mean-square amplitude of linearly-extrapolated density fluctuations σ as a function of mass M (in solar masses $M_\odot$, within a spherical top-hat filter) at different redshifts z. Halos form in regions that exceed the background density by a factor of order unity. This threshold is only surpassed by rare (many-σ) peaks for high masses at high redshifts. When discussing the abundance of halos, we will factor out the linear growth of perturbations and use the function $\sigma(M)$ at $z = 0$. The comoving radius of an unperturbed sphere containing a mass M is $R = 1.85\ \mathrm{Mpc}(M/10^{12}M_\odot)^{1/3}$.

2.2 THERMAL HISTORY DURING THE DARK AGES: COMPTON COOLING ON THE CMB

In addition to the density evolution, the second key "initial condition" for galaxy formation is the temperature of the hydrogen and helium gas that will collapse into the first galaxies. If it were isolated, the gas would simply cool adiabatically with the overall expansion of the universe. In general, this cooling rate can be written as $(\gamma-1)(\dot{\rho}_b/\rho_b)T_e$, where ρ_b is the baryon density and $\gamma = 5/3$ is the adiabatic index of a mono-atomic gas. For gas at the mean density, the factor $(\dot{\rho}/\rho) = -3H$ is just due to the Hubble expansion. However, in an overdense region, where expansion is slowed by gravity and eventually even turns into contraction, this cooling is slower (and may turn into heating); in an underdense region, the cooling accelerates. Thus, the cosmic gas (also called, "intergalactic medium" and abbreviated as IGM) will be seeded by small temperature fluctuations reflecting its density structure.

However, this is not the entire story because of the CMB. Although cosmological recombination at $z \sim 1100$ results in a nearly neutral universe, a small fraction $\sim 10^{-4}$ of electrons remain free until the era of the first galaxies. These free electrons scatter off CMB photons and bring the gas closer to equilibrium with the radiation field.

A free electron moving at a speed $v \ll c$ relative to the cosmic rest frame would probe a Doppler shifted CMB temperature with a dipole pattern,

$$T(\theta) = T_\gamma \left(1 + \frac{v}{c}\cos\theta\right), \tag{2.22}$$

where θ is the angle relative to its direction of motion and T_γ is the average CMB temperature. Naturally, the radiation will exert a friction force on the electron opposite to its direction of motion. The CMB energy density within a solid angle $d\Omega = d\cos\theta d\phi$ (in spherical coordinates) would be $d\epsilon = aT^4(\theta)d\Omega/4\pi$. Since each photon carries a momentum equal to its energy divided by c, the electron will be slowed down along its direction of motion by a net momentum flux $c(d\epsilon/c) \times \cos\theta$. The product of this momentum flux and the Thomson (Compton) cross-section of the electron (σ_T) yields the net drag force acting on the electron,

$$m_e\frac{dv}{dt} = -\int \sigma_T \cos\theta d\epsilon = -\frac{4}{3c}\sigma_T a T_\gamma^4 v. \tag{2.23}$$

The rate of energy loss by the electron is obtained by multiplying the drag force with v, giving

$$\frac{d}{dt}E = -\frac{8\sigma_T}{3m_e c}aT_\gamma^4 E, \tag{2.24}$$

where $E = \frac{1}{2}m_e v^2$. For a thermal ensemble of electrons at a non-relativistic temperature T, the average energy is $\langle E\rangle = \frac{3}{2}k_B T_e$. If the electrons reach thermal equilibrium with the CMB, then the net rate of energy exchange must vanish. Therefore, there must be a stochastic heating term which balances the above cooling term when $T = T_\gamma$. The origin of this heating term is obvious. Electrons starting at rest will be pushed around by the fluctuating electric field of the CMB

until the ensemble reaches an average kinetic energy per electron of $\langle E \rangle = \frac{3}{2} k_B T_\gamma$, at which point it stays in thermal equilibrium with the radiation.

The temperature evolution of gas at the mean cosmic density, which cools only through its coupling to the CMB and its adiabatic Hubble expansion (with no radiative cooling due to atomic transitions or heating by galaxies), is therefore described by the equation,

$$\frac{dT_e}{dt} = \frac{x}{(1+x)} \frac{8\sigma_T a T_\gamma^4}{3 m_e c} (T_\gamma - T_e) - 2HT_e, \tag{2.25}$$

where x is the fraction of all electrons which are free. For an electron-proton gas, $x = n_e/(n_e + n_H)$ where n_e and n_H are the electron and hydrogen densities, and $T_\gamma \propto (1+z)$. The second term on the right-hand-side of equation (**??**), $-2HT_e$, yields the adiabatic scaling $T_e \propto (1+z)^2$ in the absence of energy exchange with the CMB.

The relative importance of these two heating and cooling mechanisms therefore depends on the residual fraction of free electrons after cosmological recombination. The rate at which electrons recombine is $\dot{n}_e = -\alpha_B x^2 n_b^2$, where $\alpha_B \propto T_e^{-0.7}$ is the case-B recombination coefficient.[iv] Using our preferred cosmological parameters, the fractional change in x per Hubble time is therefore

$$\frac{\dot{n}_e}{H n_e} \approx 7x(1+z)^{0.8}. \tag{2.26}$$

Electrons "freeze-out" when this factor becomes of order unity; after that point, the Hubble expansion time is shorter than the recombination time. More precise numerical calculations give $x \approx 3 \times 10^{-4}$ at $z \approx 200$, as shown in Figure 2.2.

Inserting this value into equation (2.25), we find that Compton cooling becomes inefficient at $z \approx 160$. This small fraction of electrons suffices to maintain thermal contact between the baryons and CMB until $z \sim 300$, when Compton heating becomes inefficient. Figure 2.2 shows a more precise calculation: note how the gas and CMB temperatures begin to depart at $z \sim 200$, and the gas begins to follow the expected adiabatic evolution $T_e \propto (1+z)^2$ at $z \sim 100$.

Note, however, that Compton cooling can become important again if the Universe is "reionized" by stars or quasars; once $x \approx 1$, the Compton cooling time is still shorter than the age of the Universe (and hence significant relative to adiabatic cooling) down to a redshift $z \sim 6$.

[iv] This ignores recombinations to the ground state, which generate a new ionizing photon and so do not change the net ionized fraction.

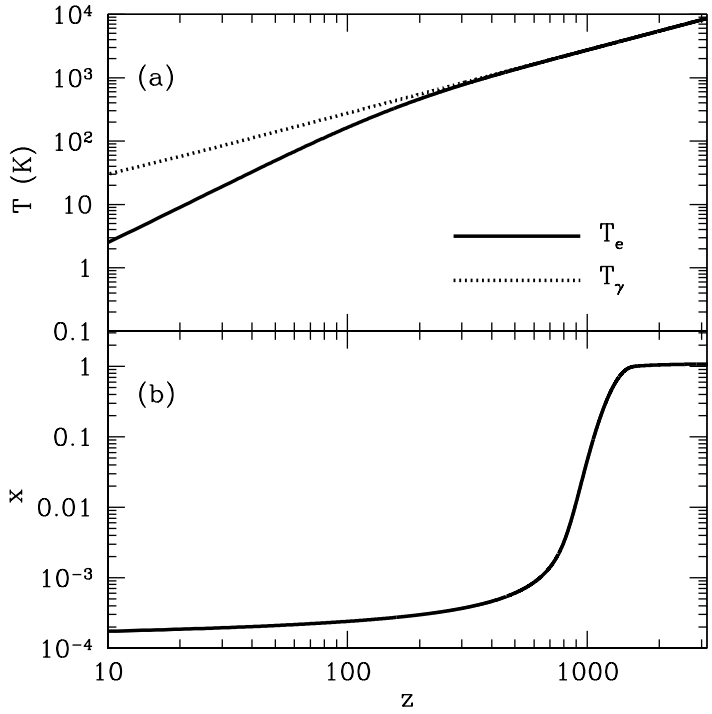

Figure 2.2 Thermal and ionization history of the Universe before the first stars form (panels a and b, respectively). In the top panel, the solid and dotted curves show T_e and T_γ, respectively. Note how the ionized fraction x decreases rapidly after recombination at $z \sim 1100$ and then "freezes-out" at $z \sim 300$. Meanwhile, Compton scattering keeps $T_e \approx T_\gamma$ until $z \sim 200$, after which the declining CMB energy density and small residual ionized fraction are no longer sufficient to maintain thermal contact between the gas and CMB. At later times, $T_e \propto (1+z)^2$ as appropriate for an adiabatically expanding non-relativistic gas. These results were produced with the publicly available code RECFAST (**http://www.astro.ubc.ca/people/scott/recfast.html**).

Chapter Three

Nonlinear Structure and Halo Formation

3.1 SPHERICAL COLLAPSE

Existing cosmological data suggests that the dark matter is "cold," that is, its pressure is negligible during the gravitational growth of galaxies. This makes the nonlinear evolution relatively simple, as it depends purely on the gravitational force. We can therefore make some progress in understanding galaxy formation by considering models for this gravitational growth that are sufficiently simple to extend into the nonlinear regime.

For simplicity, let us consider an isolated, spherically symmetric density or velocity perturbation of the smooth cosmological background and examine the dynamics of a test particle at a radius r relative to the center of symmetry. Birkhoff's theorem implies that we may ignore the mass outside this radius in computing the motion of our particle. The equation of motion describing the system reduces to the usual Friedmann equation for the evolution of the scale factor of a homogeneous Universe, but with a density parameter Ω that now takes account of the additional mass interior to the shell and its modified expansion velocity. In particular, despite the arbitrary density and velocity profiles given to the perturbation, only the total mass interior to the particle's radius and the peculiar velocity at the particle's radius contribute to the effective value of Ω. We may thus find a solution to the particle's motion which describes its departure from the background Hubble flow and its subsequent collapse or expansion. This solution holds until our particle crosses paths with one from a different radius, which happens rather late for most initial conditions.

As with the Friedmann equation for a smooth Universe, it is possible to reformulate the problem in a Newtonian form. At some early epoch corresponding to a scale factor $a_i \ll 1$, we consider a spherical patch of uniform overdensity δ_i, making a so-called 'top-hat' perturbation. If Ω_m is essentially unity at this time and if the perturbation is a pure growing mode, then the initial peculiar velocity is radially inward with magnitude $\delta_i H(t_i) r/3$, where $H(t_i)$ is the Hubble constant at the initial time and r is the radius from the center of the sphere. This can be easily derived from mass conservation (continuity equation) in spherical symmetry. The collapse of a spherical top-hat perturbation beginning at radius r_i is described by

$$\frac{d^2 r}{dt^2} = H_0^2 \Omega_\Lambda \, r - \frac{GM}{r^2} \, , \tag{3.1}$$

where r is the radius in a fixed (not comoving) coordinate frame, H_0 is the present-day Hubble constant, and the unperturbed Hubble flow velocity (to which the

above-mentioned peculiar velocity should be added) is given by $dr/dt = H(t)r$. The total mass enclosed within radius r is, $M = (4\pi/3)r_i^3\rho_i(1+\delta_i)$, with ρ_i being the background density of the Universe at time t_i. We next define the dimensionless radius $x = ra_i/r_i$ and rewrite equation (3.1) as

$$\frac{l}{H_0^2}\frac{d^2x}{dt^2} = -\frac{\Omega_m}{2x^2}(1+\delta_i) + \Omega_\Lambda x, \tag{3.2}$$

where we assume a flat universe with $\Omega_\Lambda = 1 - \Omega_m$. Our initial conditions for the integration of this orbit are

$$x(t_i) = a_i \tag{3.3}$$

$$\frac{dx}{dt}(t_i) = H(t_i)x(t_i)\left(1 - \frac{\delta_i}{3}\right) = H_0 a_i \left(1 - \frac{\delta_i}{3}\right)\sqrt{\frac{\Omega_m}{a_i^3} + \Omega_\Lambda}, \tag{3.4}$$

where $H(t_i) = H_0[\Omega_m/a_i^3 + (1-\Omega_m)]^{1/2}$ is the Hubble parameter for a flat Universe at the initial time t_i. Integrating equation (3.2) yields

$$\frac{1}{H_0^2}\left(\frac{dx}{dt}\right)^2 = \frac{\Omega_m}{x}(1+\delta_i) + \Omega_\Lambda x^2 + K, \tag{3.5}$$

where K is a constant of integration. Evaluating this at the initial time and dropping terms of order a_i (with $\delta_i \propto a_i$), we find

$$K = -\frac{5\delta_i}{3a_i}\Omega_m. \tag{3.6}$$

If K is sufficiently negative, the particle will turn-around and the sphere will collapse to zero size at a time

$$H_0 t_{coll} = 2\int_0^{a_{\max}} da\left(\Omega_m/a + K + \Omega_\Lambda a^2\right)^{-1/2}, \tag{3.7}$$

where $a_{\max}$ is the value of a which sets the denominator of the integrand to zero.

It is easier to solve the equation of motion analytically for the regime in which the cosmological constant is negligible, $\Omega_\Lambda = 0$ and $\Omega_m = 1$ (adequate for describing redshifts $1 < z < 10^3$). There are three branches of solutions: one in which the particle turns around and collapses, another in which it reaches an infinite radius with some asymptotically positive velocity, and a third intermediate case in which it reaches an infinite radius but with a velocity that approaches zero. These cases may be written as:

$$\left.\begin{array}{l} r = A(\cos\eta - 1) \\ t = B(\eta - \sin\,\eta) \end{array}\right\} \qquad \text{Closed} \qquad (0 \le \eta \le 2\pi) \tag{3.8}$$

$$\left.\begin{array}{l} r = A\eta^2/2 \\ t = B\eta^3/6 \end{array}\right\} \qquad \text{Flat} \qquad (0 \le \eta \le \infty) \tag{3.9}$$

$$\left.\begin{array}{l} r = A(\cosh\eta - 1) \\ t = B(\sinh\eta - \eta) \end{array}\right\} \qquad \text{Open} \qquad (0 \le \eta \le \infty) \tag{3.10}$$

where $A^3 = GMB^2$ applies in all cases. All three solutions have $r^3 = 9GMt^2/2$ as t goes to zero, which matches the linear theory expectation that the perturbation amplitude get smaller as one goes back in time. In the closed case, the shell turns around at time πB and radius $2A$ (when its density contrast relative to the background of an $\Omega_m = 1$ Universe is $9\pi^2/16 = 5.6$), and collapses to zero radius at time $2\pi B$. Interestingly, these collapse times are independent of the initial distance from the origin: perturbations with fixed initial density contrast collapse homologously, with all shells turning around and collapsing at the same time.

This is the fully nonlinear solution for the simplified problem of collapse of a purely spherical top hat perturbation. Of course, the real density distribution of the Universe is much more complicated. Although we cannot describe analytically the full nonlinear evolution of density perturbations, we *can* fully describe their linear evolution. A compromise is then to use this linear evolution to identify regions (such as galaxies) where spherical nonlinear evolution is not a bad approximation. It is therefore useful to determine the mapping between the *linear* density field described by perturbation theory and the *nonlinear* densities in the spherical model.

To do this, we are faced with the problem of relating the spherical collapse parameters A, B, and M to the linear theory density perturbation δ. This exercise is straightforward for the case of Ω_Λ = 0 and $\Omega_m = 1$. $K > 0$ ($K < 0$) produces an open (closed) model. Comparing coefficients in the energy equation (3.5) and the integral of the equation of motion, one finds

$$A = \frac{r_i}{2a_i}\left(\frac{5\delta_i}{3a_i}\right)^{-1} \tag{3.11}$$

$$B = \frac{1}{2H_0}\left(\frac{5\delta_i}{3a_i}\right)^{-3/2}. \tag{3.12}$$

In an $\Omega = 1$ Universe, where $1 + z = (3H_0t/2)^{-2/3}$, we find that a shell collapses at redshift $1 + z_c = 0.5929\delta_i/a_i$. Noting that, *in linear theory*, perturbations grow as $\delta \propto t^{2/3} \propto a$ in the matter dominated era, the quantity δ/a is constant with time. Thus, a shell collapsing at redshift z_c had a *linearized* overdensity extrapolated to the present day[i] of

$$\delta_{\rm crit}(z_c) = \frac{1.686}{D(z_c)} \approx 1.686(1 + z_c), \tag{3.13}$$

where $D(z)$ is the linear growth factor, although the true density (computed with the full nonlinear theory) differs. This critical density plays a key role in calculations of the halo abundance below.

Of course, we do not expect a real object to collapse to a zero size; anisotropies and angular momentum in the initial distribution will prevent perfect collapse. Instead, we envision that the material will *virialize*, with strong particle interactions transforming the bulk kinetic energy of collapse into random velocities. The result is a *dark matter halo* with a centrally-concentrated mass distribution; we will discuss the properties of such halos in §3.4 below.

[i] Linear evolution also gives $\delta_0 = 1.063(1 + z_c)$ at turnaround.

While this derivation has been for spheres of constant density, we may treat a general spherical density profile $\delta_i(r)$ up until shell crossing. A particular radial shell evolves according to the mass interior to it; therefore, we define the average overdensity $\overline{\delta_i}$

$$\overline{\delta_i}(R) = \frac{3}{4\pi R^3}\int_0^R d^3r\delta_i(r), \tag{3.14}$$

so that we may use $\overline{\delta_i}$ in place of δ_i in the above formulae. If $\overline{\delta_i}$ is not monotonically decreasing with R, then the spherical top-hat evolution of two different radii will predict that they cross each other at some late time; this is known as shell crossing and signals the breakdown of the solution. Even well-behaved $\overline{\delta_i}$ profiles will produce shell crossing if shells are allowed to collapse to $r = 0$ and then re-expand, since these expanding shells will cross infalling shells. In such a case, first-time infalling shells will never be affected prior to their turn-around; the more complicated behavior after turn-around is a manifestation of virialization. While the end state for general initial conditions cannot be predicted, various results are known for a self-similar collapse, in which $\delta(r)$ is a power-law, as well as for the case of secondary infall models.

3.2 COSMOLOGICAL JEANS MASS

Of course, the most interesting components of galaxies – stars, quasars, and people – are not made of dark matter but of baryons. As the density contrast between a spherical gas cloud and its cosmic environment grows, two main forces which come into play. The first is **gravity** and the second involves the **pressure gradient** of the gas. The second modifies the simple picture of spherical collapse above for the baryonic matter.

We can obtain a rough estimate for the relative importance of these forces from the following simple considerations. The increase in gas density near the center of the cloud sends out a pressure wave which propagates out at the speed of sound $c_s \sim (k_B T/m_p)^{1/2}$ where T is the gas temperature. The wave tries to even out the density enhancement, consistent with the tendency of pressure to resist collapse. At the same time, gravity pulls the cloud together in the opposite direction. The characteristic time-scale for the collapse of the cloud is given by its radius R divided by the free-fall speed $\sim (2GM/R)^{1/2}$, yielding $t_{\rm coll} \sim (G\langle\rho\rangle)^{-1/2}$ where $\langle\rho\rangle = M/\frac{4\pi}{3}R^3$ is the characteristic density of the cloud as it turns around on its way to collapse.[ii]

If the sound wave does not have sufficient time to traverse the cloud during the free-fall time, namely $R > R_{\rm J} \equiv c_s t_{\rm coll}$, then the cloud will collapse. Under these circumstances, the sound wave moves outward at a speed that is slower than the inward motion of the gas, and so the wave is simply carried along together with the infalling material. On the other hand, the collapse will be inhibited by pressure for

[ii] Substituting the mean density of the Earth to this expression yields the characteristic time it takes a freely-falling elevator to reach the center of the Earth from its surface ($\sim 1/3$ of an hour), as well as the order of magnitude of the time it takes a low-orbit satellite to go around the Earth (~ 1.5 hours).

a sufficiently small cloud with $R < R_{\rm J}$. The transition between these regimes is defined by the so-called Jeans radius, $R_{\rm J}$, corresponding to the Jeans mass,

$$M_{\rm J} = \frac{4\pi}{3}\langle\rho\rangle R_{\rm J}^3. \tag{3.15}$$

This mass corresponds to the total gravitating mass of the cloud, including the dark matter. As long as the gas temperature is not very different from the CMB temperature, the value of $M_{\rm J} \sim 10^5 M_\odot$ is independent of redshift.[10] This is the minimum total mass of the first gas clouds to collapse ~ 100 million years after the Big Bang.

A few hundred million years later, once the cosmic gas was ionized and heated to a temperature $T > 10^4$K by the first galaxies, the minimum galaxy mass had risen above $\sim 10^8 M_\odot$. At even later times, the UV light that filled up the Universe was able to boil the uncooled gas out of the shallowest gravitational potential wells of mini-halos with a characteristic temperature below 10^4K.[11] Below we derive the above estimates more rigorously in the cosmological context of an expanding Universe.

Similarly to the discussion above, the Jeans length $\lambda_{\rm J}$ was originally defined in Newtonian gravity as the critical wavelength that separates oscillatory and exponentially-growing density perturbations in an infinite, uniform, and stationary distribution of gas. On scales ℓ smaller than $\lambda_{\rm J}$, the sound crossing time, ℓ/c_s is shorter than the gravitational free-fall time, $(G\rho)^{-1/2}$, allowing the build-up of a pressure force that counteracts gravity. On larger scales, the pressure gradient force is too slow to react to a build-up of the attractive gravitational force. The Jeans mass is defined as the mass within a sphere of radius $\lambda_{\rm J}/2$, $M_{\rm J} = (4\pi/3)\rho(\lambda_{\rm J}/2)^3$. In a perturbation with a mass greater than $M_{\rm J}$, the self-gravity cannot be supported by the pressure gradient, and so the gas is unstable to gravitational collapse. The Newtonian derivation of the Jeans instability suffers from a conceptual inconsistency, as the unperturbed gravitational force of the uniform background must induce bulk motions. However, this inconsistency is remedied when the analysis is done in an expanding Universe.

The perturbative derivation of the Jeans instability criterion can be carried out in a cosmological setting by considering a sinusoidal perturbation superposed on a uniformly expanding background. Here, as in the Newtonian limit, there is a critical wavelength $\lambda_{\rm J}$ that separates oscillatory and growing modes. Although the expansion of the background slows down the exponential growth of the amplitude to a power-law growth, the fundamental concept of a minimum mass that can collapse at any given time remains the same.

We consider a mixture of dark matter and baryons with density parameters $\Omega_{\rm dm}(z) = \bar\rho_{\rm dm}/\rho_{\rm c}$ and $\Omega_{\rm b}(z) = \bar\rho_{\rm b}/\rho_{\rm c}$, where $\bar\rho_{\rm dm}$ is the average dark matter density, $\bar\rho_{\rm b}$ is the average baryonic density, $\rho_{\rm c}$ is the critical density, and $\Omega_{\rm dm}(z) + \Omega_{\rm b}(z) = \Omega_m(z)$. We also assume spatial fluctuations in the gas and dark matter densities with the form of a single spherical Fourier mode on a scale much smaller than the horizon,

$$\frac{\rho_{\rm dm}(R,t) - \bar\rho_{\rm dm}(t)}{\bar\rho_{\rm dm}(t)} = \delta_{\rm dm}(t)\frac{\sin(kR)}{kR}\,, \tag{3.16}$$

$$\frac{\rho_{\rm b}(R,t) - \bar\rho_{\rm b}(t)}{\bar\rho_{\rm b}(t)} = \delta_{\rm b}(t)\frac{\sin(kR)}{kR}\,, \tag{3.17}$$

where $\bar{\rho}_{\rm dm}(t)$ and $\bar{\rho}_{\rm b}(t)$ are the background densities of the dark matter and baryons, $\delta_{\rm dm}(t)$ and $\delta_{\rm b}(t)$ are the dark matter and baryon overdensity amplitudes, R is the comoving radial coordinate, and k is the comoving perturbation wavenumber.

We adopt an ideal gas equation-of-state for the baryons with an adiabatic index (or specific heat ratio) γ=5/3. Initially, at time $t = t_{\rm i}$, the gas temperature is uniform $T_{\rm b}(R, t_{\rm i})$=$T_{\rm i}$, and the perturbation amplitudes are small $\delta_{\rm dm,i}, \delta_{\rm b,i} \ll 1$. We define the region inside the first zero of $\sin(kR)/(kR)$, namely $0 < kR < \pi$, as the collapsing "object".

The Jeans mass will clearly depend on the temperature evolution of the baryons, since that determines their overall pressure. As described in §, at very high redshifts the baryon temperature traces the CMB temperature, $T_{\rm b} \propto T_\gamma \propto (1+z)$, while at $z < z_t \sim 150$ they instead cool adiabatically, $T_{\rm b} \propto \rho_b^{(\gamma-1)} \propto (1+z)^2$.

The linear evolution of a cold dark matter overdensity, $\delta_{\rm dm}(t)$ is given by

$$\ddot{\delta}_{\rm dm} + 2H\dot{\delta}_{\rm dm} = \frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) \tag{3.18}$$

whereas the evolution of the overdensity of the baryons, $\delta_{\rm b}(t)$, with the inclusion of their pressure force is described by,

$$\ddot{\delta}_{\rm b} + 2H\dot{\delta}_{\rm b} = \frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) - \frac{kT_{\rm i}}{\mu m_p}\left(\frac{k}{a}\right)^2\left(\frac{a_{\rm i}}{a}\right)^{(1+\beta_T)}\left(\delta_{\rm b} + \frac{2}{3}\beta_T[\delta_{\rm b} - \delta_{\rm b,i}]\right). \tag{3.19}$$

Here, $H(t) = \dot{a}/a$ is the Hubble parameter at a cosmological time t, $\mu = 1.22$ is the mean atomic weight of the neutral primordial gas in units of the proton mass, and the last term describes the pressure force (being roughly $c_s^2 k^2 \delta_b / a^2$). The parameter β_T describes the temperature evolution; it is 0 when the gas remains in equilibrium with the CMB and 1 in the adiabatic limit. The last term on the right hand side (in square brackets) takes into account the extra pressure gradient force in $\nabla(\rho_{\rm b} T) = (T\nabla\rho_{\rm b} + \rho_{\rm b}\nabla T)$, arising from the temperature gradient which develops in the adiabatic limit.

The Jeans wavelength $\lambda_{\rm J} = 2\pi/k_{\rm J}$ is obtained by setting the right-hand side of equation (3.19) to zero, and solving for the critical wavenumber $k_{\rm J}$. As can be seen from equation (3.19), the critical wavelength $\lambda_{\rm J}$ (and therefore the mass $M_{\rm J}$) is in general time-dependent. We infer from equation (3.19) that as time proceeds, perturbations with increasingly smaller initial wavelengths stop oscillating and start to grow.

To estimate this, we further approximate $\delta_{\rm b} \sim \delta_{\rm dm}$, and consider sufficiently high redshifts at which the Universe is matter dominated (so that $\Omega_m \approx 1$), Following cosmological recombination at $z \approx 10^3$, the residual ionization of the cosmic gas keeps its temperature locked to the CMB temperature (via Compton scattering) down to a redshift of $z_t \approx 160$ (see §2.2 and Figure 2.2) In the redshift range between recombination and z_t, $\beta_T = 0$ and

$$k_{\rm J} \equiv (2\pi/\lambda_{\rm J}) = [2kT_\gamma(0)/3\mu m_p]^{-1/2}\sqrt{\Omega_m}H_0\ , \tag{3.20}$$

so that the Jeans mass is redshift independent and obtains a value (for the total mass

of baryons and dark matter),

$$M_{\rm J} \equiv \frac{4\pi}{3}\left(\frac{\lambda_{\rm J}}{2}\right)^3 \bar{\rho}(0) = 1.35\times10^5 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} M_\odot\ . \tag{3.21}$$

At $z < z_t$, the gas temperature declines adiabatically as $[(1+z)/(1+z_t)]^2$ (i.e., $\beta_T = 1$) and the total Jeans mass obtains the value,

$$M_{\rm J} = 4.54\times10^3 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} \left(\frac{\Omega_b h^2}{0.022}\right)^{-3/5} \left(\frac{1+z}{10}\right)^{3/2} M_\odot. \tag{3.22}$$

So far, we have ignored similar effects in the dark matter component: although these collisionless particles do not feel a pressure force, their intrinsic velocity dispersion plays an analogous role to pressure, and a similar criterion for collapse exists. However, in popular *cold* dark matter models with weakly-interacting massive particles, the Jeans mass of the dark matter alone is negligible but non zero, of the order of the mass of a planet like Earth or Jupiter.[12] All halos between this minimum clump mass and $\sim 10^5 M_\odot$ are expected to contain mostly dark matter and little ordinary matter. *Warm* dark matter, with a moderately large velocity dispersion, could change this expectation and – if its Jeans mass exceeds that of the baryons – substantially modify the early phases of structure formation.

It is not clear how the value of the Jeans mass derived above relates to the mass of collapsed, bound objects. The above analysis is perturbative (equations 3.18 and 3.19 are valid only as long as $\delta_{\rm b}$ and $\delta_{\rm dm}$ are much smaller than unity), and thus can only describe the initial phase of the collapse. As $\delta_{\rm b}$ and $\delta_{\rm dm}$ grow and become larger than unity, the density profiles start to evolve and dark matter shells may cross baryonic shells due to their different dynamics. Hence the amount of mass enclosed within a given baryonic shell may increase with time, until eventually the dark matter pulls the baryons with it and causes their collapse even for objects below the Jeans mass.

Even within linear theory, the Jeans mass is related only to the evolution of perturbations at a given time. When the Jeans mass itself varies with time, the overall suppression of the growth of perturbations depends on a time-weighted Jeans mass. The proper time-weighted mass is called the filtering mass[13] $M_F = (4\pi/3)\,\bar{\rho}\,(\pi a/k_F)^3$, written in terms of the comoving wavenumber k_F associated with the "filtering scale". This scale can be derived as follows.

Consider a growing mode perturbation in the dark matter $\delta_{\rm dm}$ and baryons $\delta_{\rm b}$ in the limit where the baryons are gravitationally unimportant ($\Omega_b \ll \Omega_m$). In this regime, the linear perturbation equations admit a simple solution in the special case where the Jeans wavenumber $k_{\rm J}$ is constant in time,

$$\delta_{\rm b}(t,k) = \frac{\delta_{\rm dm}(t,k)}{1+k^2/k_{\rm J}^2}, \tag{3.23}$$

where the dark matter fluctuation grows in proportion to the linear growth factor, $\delta_{\rm dm} \propto D(t)$. In the general case where the Jeans wavenumber $k_{\rm J}$ is time dependent, we can identify the proper time averaging by considering the perturbative effect of gas pressure on large scales. We therefore expand the ratio $\delta_{\rm b}(t,k)/\delta_{\rm dm}(t,k)$ in powers of k^2 with $\delta_{\rm b}(t,k=0) = \delta_{\rm dm}(t,k=0)$. The ratio between the linear

overdensity of the baryons and dark matter in the limit of small k can then be written as

$$\frac{\delta_{\rm b}}{\delta_{\rm dm}} = 1 - \frac{k^2}{k_F^2} + ... \ . \tag{3.24}$$

or equivalently

$$\frac{\delta_{\rm b}(t,k)}{\delta_{\rm dm}(t,k)} = 1 - \frac{A(t)}{D(t)}k^2, \tag{3.25}$$

where $A(t) \equiv D(t)/k_F^2$ can be solved for by substituting the latter relation into the coupled linear growth equations for $delta_{\rm b}$ and $\delta_{\rm dm}$ and ignoring terms of order k^4 or higher. This gives the differential equation,

$$\frac{d^2A}{dt^2} + 2H\frac{dA}{dt} = \frac{c_s^2}{a^2}D(t), \tag{3.26}$$

where c_s is the sound speed of the baryons. The solution to this equation gives the filtering wavenumber k_F is terms of time integrals of the Jeans wavenumber $k_{\rm J}$ (using the latter's relation to c_s),

$$\frac{1}{k_F^2(t)} = \frac{1}{D(t)}\int_0^t dt'\, a^2(t')\frac{\ddot{D}(t') + 2H(t')\dot{D}(t')}{k_{\rm J}^2(t')}\int_{t'}^t \frac{dt''}{a^2(t'')} \ . \tag{3.27}$$

At high redshifts (where $\Omega_m(z) \rightarrow 1$), this relation simplifies to

$$\frac{1}{k_F^2(t)} = \frac{3}{a}\int_0^a \frac{da'}{k_{\rm J}^2(a')}\left(1 - \sqrt{\frac{a'}{a}}\right) \ . \tag{3.28}$$

It is conventional to assume that the Jeans or filtering mass accurately reflects the threshold for baryonic structure formation. However, linear theory specifies whether an initial perturbation, characterized by the parameters k, $\delta_{\rm dm,i}$, $\delta_{\rm b,i}$ and $t_{\rm i}$, begins to grow. To determine the minimum mass of the resulting nonlinear baryonic object following the shell-crossing and virialization of the dark matter, we typically appeal to the spherical collapse model described in the previous section.

3.3 PRIMORDIAL STREAMING OF BARYONS RELATIVE TO DARK MATTER

Prior to cosmological recombination, the baryons and the cosmic background radiation were tightly coupled and behaved as a single fluid, separate from the dark matter. The primordial density fluctuations produced acoustic waves in the radiation-baryon fluid. When the gas decoupled from the radiation at $z \approx 10^3$, it was streaming relative to the dark matter with a root-mean-square (*rms*) speed of $v_{\rm bc} \approx 10^{-4}c = 30 \ {\rm km\ s^{-1}}$. Figure 3.1 shows the variance of the velocity difference perturbations (in units of c) per $\ln k$ as a function of the mode wavenumber k at $z = 10^3$. The power extends to scales as large as the sound horizon at recombination, ~ 140 comoving Mpc, but declines rapidly at $k > 0.5 \ {\rm Mpc}^{-1}$, indicating that the velocity of the baryons relative to the dark matter was coherent over the photon

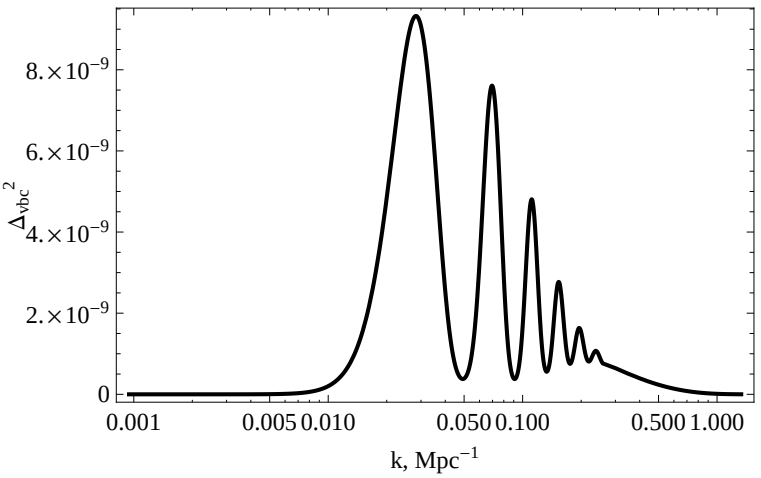

Figure 3.1 The variance of the velocity difference perturbations (in units of c) between baryons and dark matter per $\ln k$ as a function of comoving wavenumber k at $z = 10^3$. Figure credit: D. Tseliakhovic & C. Hirata, Phys. Rev. **D82**, 3520 (2010).

diffusion (Silk damping) scale of several comoving Mpc. This scale is larger by two orders of magnitude than the size of the regions out of which the first galaxies were assembled at later times. Therefore, in the rest-frame of those galaxies, the background intergalactic baryons appeared to be moving coherently as a wind. It is therefore interesting to examine whether this wind had a significant effect on the assembly of baryons onto the earliest galaxies.

Following recombination, the neutral gas was freed to fall into the gravitational potential wells of the dark matter and so its velocity difference from the dark matter declined as $\mathbf{v}_{\rm bc} \propto (1+z)$. This implies that by $z \sim 50$, the typical streaming velocity $\sim (50/10^3) \times 30 = 1.5\ {\rm km\ s^{-1}}$ corresponded to an equivalent gas temperature $T_{\rm bc} \sim m_p v_{\rm bc}^2/k_B = 270\ {\rm K}[(1+z)/50]^2$, comparable to the virial temperature of the first gas clouds that cooled through molecular hydrogen (H_2) transitions. Therefore, in the frame of the dark matter the baryonic wind have increased by a factor of order unity the minimum halo mass in which the very first generation of stars could have formed. The effect was more dramatic for the filtering mass (the time-averaged Jeans mass, without the cooling constraint), which increased at $z \sim$ 20–100 from $\sim 2 \times 10^4 M_\odot$ without streaming to $\sim 2 \times 10^5 M_\odot$ at the *rms* streaming speed. Figure 3.2 shows the increase in the minimum halo mass in which gas was able to assemble and cool at various redshifts, as a function of the initial streaming velocity, $v_{\rm stream} \equiv v_{\rm bc}(z = 100)$. The *rms* value of $v_{\rm stream} = 3\ {\rm km\ s^{-1}}$ at $z = 100$ is marked by the horizontal line. As expected, the infall of gas into halos more massive than $\sim 10^6 M_\odot$ for which the virial temperature $T_{\rm vir} \gg T_{\rm bc}$, was not affected significantly by the baryonic wind.

In linear perturbation theory (§2), each Fourier mode evolves independently.

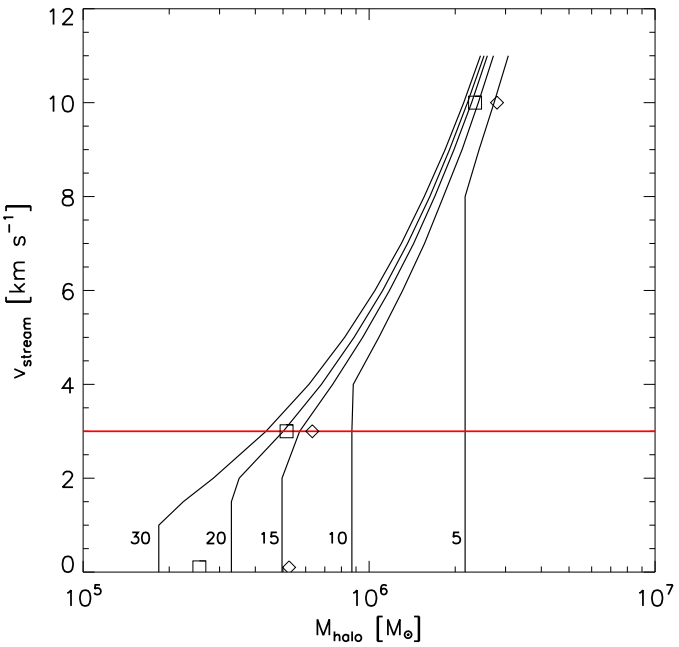

Figure 3.2 Effect of initial streaming speed $v_{\rm stream}$ at $z = 100$ on the minimum halo mass into which gas can assemble later and form stars (with the horizontal bar at $3\ {\rm km\ s^{-1}}$ marking the expected *rms* value). Each line represents the necessary halo mass for baryon collapse at the labeled redshift. Diamonds represent the final halo masses found in standard collapse simulations ($z = 14$ with no streaming), squares represent masses from accelerated collapse simulations ($z = 24$ with no streaming), and the lines delineate the prediction of a simple analytic model. The halo masses do not increase significantly at low streaming velocities. Halos collapsing at high redshift are more affected by relative streaming, as the physical streaming velocities are higher at these early times. Figure credit: A. Stacy, V. Bromm, & A. Loeb, MNRAS, in press (2010); arXiv:1011.4512.

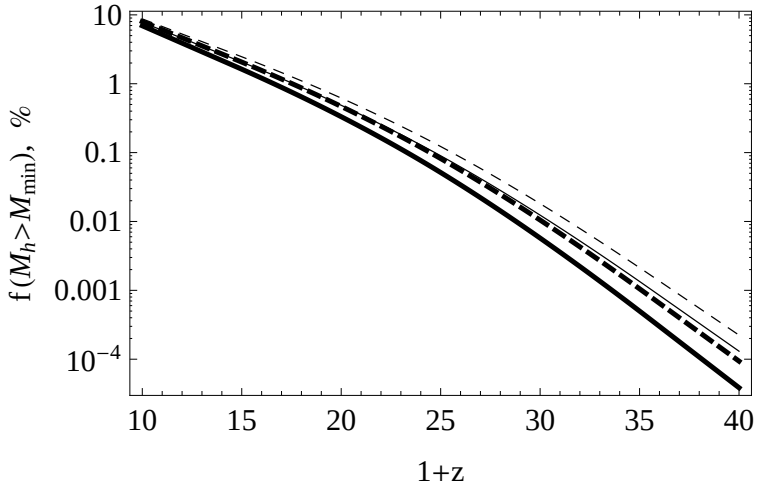

Figure 3.3 The total mass fraction (in percent) in halos above the filtering mass (dashed lines) and above the H_2 cooling mass (solid lines). Thin lines ignore baryon streaming and thick lines include it. Figure credit: D. Tseliakhovic, R. Barkana, & C. Hirata, MNRAS, submitted (2010); arXiv:1012.2574.

However, the streaming effect on the growth of structure[14] is manifestly non-linear and results from second-order terms, such as the convective derivative $(\mathbf{v_k}\cdot\mathbf{k})\mathbf{v_k}$ in the equation of motion, which couple small-scale (high-k) modes – as they become nonlinear, to large scale (low-k) modes. Effectively, the relative velocity acts as an increased sound speed (since it needs to be dissipated upon virialization of the gas) and prevents the baryons from settling into the shallowest potential wells. Figure 3.3 compares the fraction of baryons that collapsed into halos above the H_2 cooling threshold with and without streaming.

Since the streaming velocity varied spatially, the formation of the earliest Population III stars was modulated on large spatial scales of up to ~ 140 cMpc. The effects of baryonic streaming were most pronounced in the lowest mass halos ($< 10^6 M_\odot$) and at the highest redshifts ($z > 20$) – when the global collapse fraction of baryons and the corresponding radiative effects of stars were small.

The baryonic streaming alters the growth of structure, causing a small ($\sim 15\%$) suppression in the matter power spectrum around the Jeans wavenumber $k_{\rm J} \sim 200{\rm Mpc}^{-1}$, and a strong scale-dependent bias for the earliest gas clouds on scales of up to ~ 140 cMpc. As long as H_2 was not quickly dissociated by the UV background produced by the first stars, the streaming effects might have left observable signatures on the 21-cm signal at redshifts $z > 20$. However, at the reionization redshifts during which the global star formation rate was dominated by halos with a virial temperature above the cooling threshold of atomic hydrogen $T_{\rm vir} > 10^4$ K and a corresponding mass $> 10^8 M_\odot[(1+z)/10]^{-3/2}$ (equation 3.33), the radiative signatures of the primordial baryonic streaming were likely negligible.

3.4 HALO PROPERTIES

When an object above the Jeans mass collapses, the dark matter forms a halo inside of which the gas may cool, condense to the center, and eventually fragment into stars. The dark matter cannot cool since it has very weak interactions. As a result, a galaxy emerges with a central core that is occupied by stars and cold gas and is surrounded by an extended halo of invisible dark matter. Since cooling eliminates the pressure support from the gas, the only force that can prevent the gas from sinking all the way to the center and ending up in a black hole is the centrifugal force associated with its rotation around the center (angular momentum). The slight ($\sim 5\%$) rotation, given to the gas by tidal torques from nearby galaxies as it turns around from the initial cosmic expansion and gets assembled into the object, is sufficient to stop its infall on a scale which is *an order of magnitude smaller* than the size of the dark matter halo[15] (the so-called "virial radius"). On this stopping scale, the gas is assembled into a thin disk and orbits around the center for an extended period of time, during which it tends to break into dense clouds which fragment further into denser clumps. Within the compact clumps that are produced, the gas density is sufficiently high and the gas temperature is sufficiently low for the Jeans mass to be of order the mass of a star. As a result, the clumps fragment into stars and a galaxy is born.

In the popular cosmological model, small objects formed first. The very first stars must have therefore formed inside gas condensations just above the cosmological Jeans mass, $\sim 10^5 M_\odot$. Whereas each of these first gaseous halos was not massive or cold enough to make more than a single high-mass star, star clusters started to form shortly afterwards inside bigger halos.

By solving the equation of motion (3.1) for a spherical overdense region, we can relate the characteristic radius and gravitational potential well of each of these galaxies to their mass and their redshift of formation. In principle, a spherical region would collapse to a point mass, but of course the real world is not so idealized. As already mentioned, even a slight violation of the exact symmetry of the initial perturbation can prevent the top-hat from collapsing to a point. Instead, the halo reaches a state of virial equilibrium through violent dynamical relaxation. We are familiar with the fact that the circular orbit of the Earth around the Sun has a kinetic energy which is half the magnitude of the gravitational potential energy. According to the *virial theorem*, this happens to be a property shared by all dynamically relaxed, self-gravitating systems. We may therefore use $U = -2K$ to relate the potential energy U to the kinetic energy K in the final state of a collapsed halo. This implies that the virial radius is half the turnaround radius (where the kinetic energy vanishes). Using this result, the final mean overdensity relative to ρ_c at the collapse redshift turns out to be $\Delta_c = 18\pi^2 \simeq 178$ in the $\Omega_m = 1$ case, which applies at redshifts $z \gg 1$. Note that the virial overdensity at collapse implies that the dynamical time within the virial radius of galaxies, $\sim (G\rho_{\rm vir})^{-1/2}$, is of order a tenth of the age of the Universe at any redshift. In a Universe with $\Omega_m + \Omega_\Lambda = 1$ the virial overdensity at the collapse redshift admits the fitting formula[16]

$$\Delta_c = 18\pi^2 + 82d - 39d^2 \,, \tag{3.29}$$

where $d \equiv \Omega_m(z) - 1$ is evaluated at the collapse redshift, so that

$$\Omega_m(z) = \frac{\Omega_m(1+z)^3}{\Omega_m(1+z)^3 + \Omega_\Lambda + \Omega_k(1+z)^2} \, . \tag{3.30}$$

A halo of mass M collapsing at redshift $z \gg 1$ thus has a virial radius

$$r_{\rm vir} = 1.5 \left[\frac{\Omega_m}{\Omega_m(z)} \, \frac{\Delta_c}{18\pi^2}\right]^{-1/3} \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{-1} \, {\rm kpc} \, , \tag{3.31}$$

and a corresponding circular velocity,

$$V_c = \left(\frac{GM}{r_{\rm vir}}\right)^{1/2} = 17.0 \left[\frac{\Omega_m}{\Omega_m(z)} \, \frac{\Delta_c}{18\pi^2}\right]^{1/6} \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{1/2} \, {\rm km \, s^{-1}} \, . \tag{3.32}$$

We may also define a virial temperature

$$T_{\rm vir} = \frac{\mu m_p V_c^2}{2k} = 1.04{\times}10^4 \, \left(\frac{\mu}{0.6}\right) \left[\frac{\Omega_m}{\Omega_m(z)} \, \frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{M}{10^8 M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \, {\rm K} \, , \tag{3.33}$$

where μ is the mean molecular weight and m_p is the proton mass. Note that the value of μ depends on the ionization fraction of the gas; for a fully ionized primordial gas $\mu = 0.59$, while a gas with ionized hydrogen but only singly-ionized helium has $\mu = 0.61$. The binding energy of the halo is approximately,

$$E_b = \frac{1}{2}\frac{GM^2}{r_{\rm vir}} = 2.9 \times 10^{53} \left[\frac{\Omega_m}{\Omega_m(z)} \, \frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{M}{10^8 M_\odot}\right)^{5/3} \left(\frac{1+z}{10}\right) \, {\rm erg} \, . \tag{3.34}$$

Note that if the ordinary matter traces the dark matter, its total binding energy is smaller than E_b by a factor of Ω_b/Ω_m, and could be lower than the energy output of a single supernova[iii] ($\sim 10^{51}$ ergs) for the first generation of dwarf galaxies.

Although spherical collapse captures some of the physics governing the formation of halos, structure formation in cold dark matter models proceeds hierarchically. At early times, most of the dark matter was in low-mass halos, and these halos then continuously accreted and merged to form high-mass halos. Numerical simulations of hierarchical halo formation indicate a roughly universal spherically-averaged density profile for the resulting halos, though with considerable scatter among different halos. This so-called NFW profile has the form[iv]

$$\rho(r) = \frac{3H_0^2}{8\pi G}(1+z)^3 \frac{\Omega_m}{\Omega_m(z)} \frac{\delta_c}{c_{\rm N} x (1 + c_{\rm N} x)^2} \, , \tag{3.35}$$

where $x = r/r_{\rm vir}$, and the characteristic density δ_c is related to the concentration parameter $c_{\rm N}$ by

$$\delta_c = \frac{\Delta_c}{3} \frac{c_{\rm N}^3}{\ln(1+c_{\rm N}) - c_{\rm N}/(1+c_{\rm N})} \, . \tag{3.36}$$

[iii] A supernova is the explosion that follows the death of a massive star.

[iv] This functional form is commonly labeled as the 'NFW profile' after the original paper by Navarro, J. F., Frenk, C. S. & White, S. D. M. *Astrophys. J.* **490**, 493 (1997).

The concentration parameter itself depends on the halo mass M, at a given redshift z, with a value of order ~ 4 for newly collapsed halos and a larger value < 20 at later times. An even better fit to state-of-the-art CDM simulations is obtained with the Einasto profile[17],

$$\ln\left[\frac{\rho(r)}{\rho_{-2}}\right] = -\frac{2}{\alpha}\left[\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right], \tag{3.37}$$

where $\alpha \approx 0.16$, r_{-2} is the radius where the logarithmic slope of the density profile equals the isothermal sphere value, $(d\ln\rho/d\ln r) = -2$. At this radius $r^2\rho$ peaks at a density value of $\rho_{-2} = \rho(r_{-2})$. For the NFW profile, $r_{-2} = r_{\rm vir}/c_{\rm N}$.

3.5 ABUNDANCE OF DARK MATTER HALOS

In addition to characterizing the properties of individual halos, a critical prediction of any theory of structure formation is the abundance of halos, namely, the number density of halos as a function of mass, at any redshift. This prediction is an important step toward inferring the abundances of galaxies and galaxy clusters. While the number density of halos can be measured for particular cosmologies in numerical simulations, an analytic model helps us gain physical understanding and can be used to explore the dependence of abundances on all the cosmological parameters.

A simple analytic model which successfully matches most of the numerical simulations was developed by Bill Press and Paul Schechter in 1974.[18] The model is based on the ideas of a Gaussian random field of density perturbations, linear gravitational growth, and spherical collapse. Once a region on the mass scale of interest reaches the threshold amplitude for collapse according to linear theory, it can be declared as a virialized object. Counting the number of such density peaks per unit volume is straightforward for a Gaussian probability distribution.

To determine the abundance of halos at a redshift z, we use δ_M, the density field smoothed on a mass scale M, as defined in §2.1.1. Since δ_M is distributed as a Gaussian variable with zero mean and a standard deviation $\sigma(M)$ (which depends only on the present linear power spectrum; see equation 2.21), the probability that δ_M is greater than some fixed δ equals

$$\int_{\delta}^{\infty} d\delta_M \frac{1}{\sqrt{2\pi}\,\sigma(M)} \exp\left[-\frac{\delta_M^2}{2\,\sigma^2(M)}\right] = \frac{1}{2}{\rm erfc}\left(\frac{\delta_M}{\sqrt{2}\,\sigma(M)}\right) . \tag{3.38}$$

The basic ansatz is to identify this probability with the fraction of dark matter particles that are part of collapsed halos of mass *greater than* M at redshift z. Note that a given region smoothed on mass M could be part of an even larger overdensity above the threshold, which is why we have the fraction of particles in halos above this mass threshold.

We need two additional ingredients to complete the model. First, we set the threshold density to $\delta_{\rm crit}(z)$ (see equation 3.13), which is the critical density of collapse found for a spherical top-hat. Crucially, $\delta_{\rm crit}$ is the *linearized* density associated with collapse in this nonlinear model, so it is directly comparable to the linearized treatment of the density field in the Gaussian approximation.

A technical note is now needed: $\delta_{\rm crit}$ is conventionally extrapolated to the present day, since $\sigma(M)$ is typically calculated using the power spectrum evaluated at $z = 0$. This is a subtle point: for $\Omega_m \approx 1$ the spherical collapse model finds that collapse occurs at a fixed fractional overdensity ($\delta = 1.686$) regardless of redshift. However, because the model makes use of linearized densities, we can evolve this threshold (at each redshift) to the present day and compare to a single linearized density field for all redshifts. In that case, the threshold for halo formation decreases with cosmic time according to equation 3.13: regions with large linearized, present-day densities collapse first, while those with more modest overdensities collapse later. Because the linear evolution is independent of scale, this also means that dense regions in the initial conditions collapse first as naturally expected.

The second key ingredient is to note that even regions with $\delta_M < 0$ can actually be part of collapsed objects, if they are part of a regions with $\delta > \delta_{\rm crit}$ on a scale $M' > M$. The original Press & Schechter paper solved this in an ad hoc fashion by multiplying the collapsed fraction of matter by a factor of two; this guarantees that every dark matter particle is part of a halo (of some $M > 0$) even if its immediate environment is underdense. Thus, the final formula for the mass fraction in halos above M at redshift z, or the *collapse fraction* is

$$f_{\rm coll}(> M|z) = {\rm erfc}\left(\frac{\delta_{\rm crit}(z)}{\sqrt{2}\,\sigma(M)}\right) . \tag{3.39}$$

We will revisit the ad-hoc factor of 2, and provide a more satisfying explanation for the adjustment in the following sections.

Differentiating the fraction of dark matter in halos above mass M yields the mass distribution. Letting dn be the comoving number density of halos of mass between M and $M + dM$, we have

$$\frac{dn}{dM} = \sqrt{\frac{2}{\pi}}\,\frac{\rho_m}{M}\,\frac{-d(\ln\sigma)}{dM}\,\nu_c\,e^{-\nu_c^2/2} , \tag{3.40}$$

where $\nu_c = \delta_{\rm crit}(z)/\sigma(M)$ is the number of standard deviations away from zero that the critical collapse overdensity represents on mass scale M. Thus, the abundance of halos depends on the two functions $\sigma(M)$ and $\delta_{\rm crit}(z)$, each of which depends on cosmological parameters.

3.5.1 The Excursion Set Formalism

Although the original Press-Schechter model is founded on an important physical insight, it turns out to be profitable to rephrase the problem in an entirely different way. This provides two benefits: first, it yields a much more satisfying derivation of the factor of two correction that is necessary, and second, it provides a number of new insights into the spatial distribution and histories of dark matter halos.

In particular, the Press-Schechter formalism makes no attempt to deal with the correlations among halos or between different mass scales. This means that, while it can generate a distribution of halos at two different epochs, it says nothing about how particular halos in one epoch are related to those in the second. We therefore

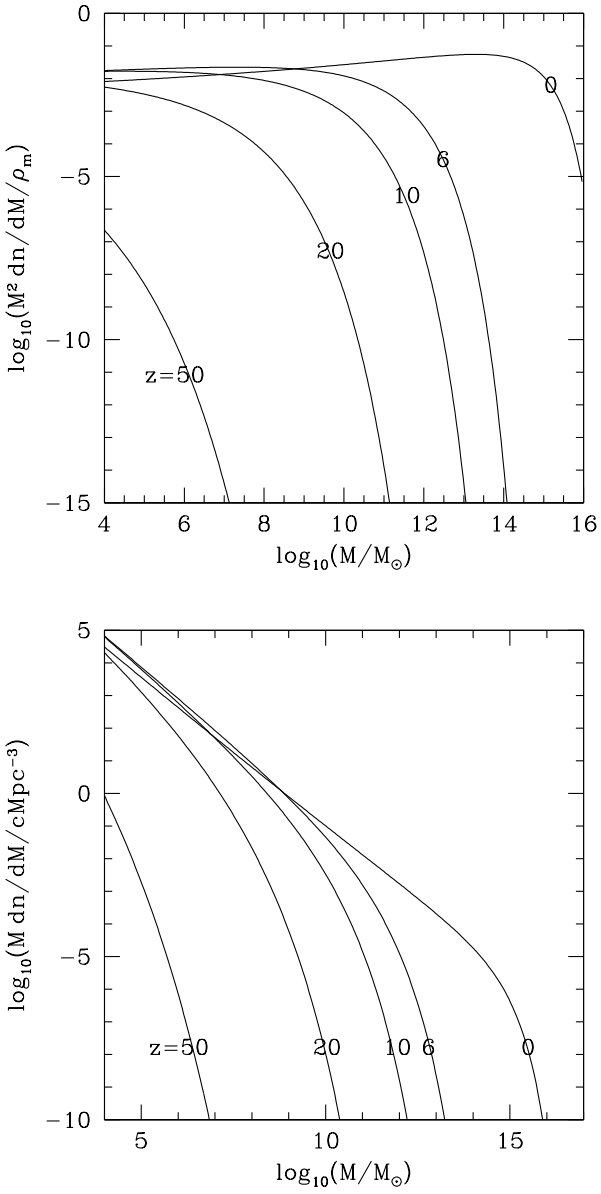

Figure 3.4 *Top:* The mass fraction incorporated into halos per logarithmic bin of halo mass $(M^2 dn/dM)/\rho_m$, as a function of M at different redshifts z. Here $\rho_m = \Omega_m \rho_c$ is the present-day matter density, and $n(M)dM$ is the comoving density of halos with masses between M and $M+dM$. The halo mass distribution was calculated based on an improved version of the Press-Schechter formalism for ellipsoidal collapse [Sheth, R. K., & Tormen, G. *Mon. Not. R. Astron. Soc.* **329**, 61 (2002)] that fits better numerical simulations. *Bottom:* Number density of halos per logarithmic bin of halo mass, Mdn/dM (in units of comoving Mpc^{-3}), at various redshifts.

would like some method to predict, at least statistically, the growth of individual halos via accretion and mergers. Even restricting ourselves to spherical collapse, such a model must utilize the full spherically-averaged density profile around a particular point. The potential correlations between the mean overdensities at different radii make the statistical description substantially more difficult.

The excursion set formalism seeks to describe the statistics of halos by considering the statistical properties of δ_M, the average overdensity within some spherical window of characteristic mass M, as a function of M (or equivalently R). While the Press-Schechter model depends only on the Gaussian distribution of $\overline{\delta}$ for one particular M, the excursion set considers all M as a set. Again the connection between a value of the linear regime δ_M and the final state is made via the spherical collapse solution so that there is a critical value $\delta_{\rm crit}(z)$ of δ_M which is required for collapse at a redshift z.

The basic idea is to view the density field around a given point, smoothed on different scales, as a diffusion process. Smoothed over a sufficiently large mass, $\delta_M \to 0$. As we zoom in to smaller scales, we naturally expect δ_M to deviate from zero, with a variance that must equal $\sigma^2(M)$. It is most natural to view this process in Fourier space: as we approach smaller scales, more and more Fourier modes become important, adding fluctuations to the density field. The particular set of modes at our point will determine the "trajectory" of δ_M as a function of smoothed mass. The key insight of the excursion set approach is that we can consider this trajectory as a diffusion process (because each k-mode is independent from all others) and thereby calculate its statistics. Conceptually, each set of Fourier modes that one adds as M decreases provides a step in the random walk of the density field, so the key is in generating the distribution of these random walks.

The subtlety in this approach lies in defining the smoothed density field; recall that it is the full (linearized) density field convolve with a window function $W(R)$. For most choices of window function, the functions δ_M are correlated from one M to another such that it is prohibitively difficult to calculate the desired statistics directly. However, for one particular choice of a window function, the correlations between different M greatly simplify and many interesting quantities may be calculated.[19] We take advantage of the fact that, in linear theory, each Fourier mode evolves independently, with no correlations between different scales k, and we use a k-space top-hat window function, namely, $W_k = 1$ for all k less than some critical k_c and $W_k = 0$ for $k > k_c$. In that case, each step in the random walk corresponds to increasing k_c. For this filter,

$$\delta_M = \int_{k<k_c(M)} \frac{d^3k}{(2\pi)^3}\delta_k, \tag{3.41}$$

meaning that the overdensity on a particular scale is simply the sum of the random variables (each Gaussian distributed) δ_k interior to the chosen k_c. Because the filter is sharp, we simply add new Fourier modes to change scales. Because these are independent of the larger-wavelength modes already inside the filter, the difference between δ_M on one mass scale and that on another mass scale is statistically independent from the value on the larger mass scale: i.e., each "step" in the walk is uncorrelated with previous steps, and the difference between the δ_M on two mass

scales is just the sum of the δ_k in the spherical k-shell between the two k_c, which is independent of the sum of the δ_k interior to the smaller k_c. We thus have a simple random walk, albeit one where the step-size varies with k_c.

Meanwhile, the distribution of δ_M (given no prior information about the random walk at larger M) is still a Gaussian of zero mean and a variance of $\sigma^2(M)$ (see equation 2.20).

Unfortunately, this filter is fundamentally inconsistent with the threshold $\delta_{\rm crit}$. The k-space top-hat filter has a *spatial* form $W(r) \propto j_1(k_c r)/k_c r$, where $j_1(x)$ is the first spherical Bessel function (see equation 2.20).[v] Thus, in real space, this set of modes exhibits a (decaying) sinusoidal oscillation rather than the sharp real-space top-hat used to derive $\delta_{\rm crit}$. Thus, at least in principle, we cannot hope to have the simultaneous advantages of real-space top-hats (specifically, the simple spherical collapse criterion) and k-space top-hats (uncorrelated steps in the random walk). Nevertheless, we may brush this inconsistency aside assuming that the two different filters are close enough to be compatible. The only justification for such an approach is its eventual success and its simplicity: although more self-consistent approaches are possible, they fare no better in the end.

It is now easy to re-derive the Press-Schechter mass function, including the previously unexplained factor of two. The fraction of mass elements included in halos of mass less than M is just the probability that a random walk remains below $\delta_{\rm crit}(z)$ for *all* k_c less than K_c, the filter cutoff appropriate to M. This probability must be the complement of the sum of the probabilities that *(a)* $\delta_M > \delta_{\rm crit}(z)$ and that *(b)* $\delta_M < \delta_{\rm crit}(z)$ but $\delta_{M'} > \delta_{\rm crit}(z)$ for some $M' > M$. The first is immediately obvious: since the distribution of δ_M is simply Gaussian with variance $\sigma^2(M)$, the fraction of random walks falling into class *(a)* is simply

$$p_a = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\delta_{\rm crit}(z)}^{\infty} d\delta \ \exp\{-\delta^2/2\sigma^2(M)\}. \tag{3.42}$$

The second class can also be easily computed. In fact, these two cases have an equal probability: any random walk belonging to class *(a)* may be reflected around its first upcrossing of $\delta_{\rm crit}(z)$ to produce a walk of class *(b)*, and vice versa. Hence, the fraction of mass elements included in halos of mass less than M at redshift z is simply

$$f_{\rm coll}(< M) = 1 - 2p_a, \tag{3.43}$$

which may be differentiated to yield the Press-Schechter mass function, equation 3.40. This approach better shows the physical origin of the extra factor of 2: many of the mass elements may *appear* to be in local underdensities but have actually already been incorporated into larger collapsed halos.

3.5.2 The Extended Press-Schechter Formalism: Conditional Mass Functions and Accretion Histories

The other advantage of the excursion set approach is that it allows us to examine how halos relate to each other. First, consider how halos at one redshift are related

[v] This implies a comoving volume $6\pi^2/k_c^3$ or mass $6\pi^2\rho_m/k_c^3$. The characteristic radius of the filter is $\sim k_c^{-1}$, as expected.

to those at another. Suppose that a halo of mass M_2 exists at redshift z_2. Then we know that the random function δ_M for each mass element within the halo first crosses $\delta_{\rm crit}(z_2)$ at k_{c2} corresponding to M_2.

Given this constraint, we may study the distribution of k_c where the function δ_M crosses other thresholds (keep in mind that $\delta_{\rm crit}$ is a function of redshift, so these other thresholds tell us about either the progenitors or "descendants" of the given halo). It is particularly easy to construct the probability distribution for when trajectories first cross some $\delta_{\rm crit}(z_1) > \delta_{\rm crit}(z_2)$ (implying $z_1 > z_2$, or the halo progenitors); clearly this must occur at some $k_{c1} > k_{c2}$ or $M_1 < M_2$.

Fortunately, this problem reduces to the previous one if we simply translate the origin of the random walks from $(k_c, \delta_M) = (0, 0)$ to $(k_{c2}, \delta_{\rm crit}[z_2])$. We therefore compute the distribution of halo masses M_1 that a mass element finds itself in at redshift z_1, given that it is part of a larger halo of mass M_2 at a later redshift z_2:

$$\frac{dP}{dM_1}(M_1, z_1|M_2, z_2) = \sqrt{\frac{2}{\pi}} \frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{[\sigma^2(M_1) - \sigma^2(M_2)]^{3/2}} \left|\frac{d\sigma(M_1)}{dM_1}\right| \exp\left\{-\frac{[\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)]^2}{2[\sigma^2(M_1) - \sigma^2(M_2)]}\right\}. \tag{3.44}$$

This may be rewritten as saying that the quantity

$$\tilde{v} = \frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{\sqrt{\sigma^2(M_1) - \sigma^2(M_2)}} \tag{3.45}$$

is distributed as the positive half of a Gaussian with unit variance; equation (3.45) may be inverted to find $M_1(\tilde{v})$.

We can interpret the statistics of these random walks as those of merging and accreting halos. For a single halo, we may imagine that as we go back in time, the object breaks into ever smaller pieces, similar to the branching of a tree. Equation (3.44) provides the distribution of the sizes of these branches at some given earlier time. One can then imagine using this description of the ensemble distribution to generate random realizations of the merger histories of single halos – or "merger trees." One recursively steps back in time, at each step breaking the final object into two pieces, choosing a value from the distribution of equation 3.44 to determine the mass ratio of the two branches.

Unfortunately, this has proven to be difficult in practice. Part of the problem is conceptual: one might want to define "merger rates" by taking the limit of equation (3.44) as $z_2 \to z_1$. However, one immediately finds that the resulting rate is not symmetric in the Press-Schechter theory: the rate at which objects of mass M merge with objects of mass M' is not equal to the rate at which objects of mass M' merge with objects of mass M! The root of the problem is that, even with the excursion set approach, the Press-Schechter formalism does not divide dark matter particles into discrete objects; rather it simply computes the statistical properties of the ensemble. Unfortunately, quantities like the merger rate implicitly assume that the objects do sit in discrete objects and ignore smooth accretion of diffuse matter. N-body simulations are the most reliable tool for following the merger statistics.

Nevertheless, we may use the distribution of the ensemble to derive some approximate analytic results that at least provide a helpful guide. A useful example

is the distribution of the epoch at which an object that has mass M_2 at redshift z_2 has accumulated half of its mass. The probability that the formation time is earlier than z_1 can be defined as the probability that at redshift z_1 a progenitor whose mass exceeds $M_2/2$ exists:

$$P(z_f > z_1) = \int_{M_2/2}^{M_2} \frac{M_2}{M} \frac{dP}{dM}(M, z_1|M_2, z_2) dM, \tag{3.46}$$

where dP/dM is given in equation (3.44). The factor of M_2/M corrects the counting from mass-weighted to number-weighting; each halo of mass M_2 can have only one progenitor of mass greater than $M_2/2$. Differentiating equation (3.46) with respect to time gives the distribution of formation times. Overall, the excursion set formalism provides a reasonable approximation to more exact numerical simulations of halo assembly and merging histories.

3.5.3 Improvements to the Press-Schechter Formalism

The above simple ansatz was refined over the years to provide a better match to numerical simulation. In particular, the Press-Schechter mass function substantially underestimates the abundance of the rare massive halos (especially at high redshift) and overestimates the abundance of low-mass halos.

There are two key areas in which the Press-Schechter approach can clearly be improved. The first is the mismatch in filter choice between the random walks and the spherical collapse model. However, more self-consistent ab initio approaches do not significantly improve the results, even at the cost of significantly increased complexity.

The second approach is to refine the spherical collapse model itself: as we will see in the next chapter, dark matter structures rarely collapse symmetrically, and so it is possible to improve the threshold density $\delta_{\rm crit}(z)$ by including a more accurate physical description. The best motivated such approach is to allow ellipsoidal collapse, in which the three axes of the object collapse at different times. The torques driving this collapse are set by the halo's environment, which depends on the halo mass itself (as will be shown in §3.6). This means that the collapse threshold $\delta_{\rm crit}^{\rm ST}$ is a function of not only redshift but also halo mass, and so the absorbing barrier in the diffusion problem is no longer a constant. In particular, the threshold increases as halos get smaller: this increases the abundance of massive halos and decreases the abundance of small halos, just as needed.

However, the match to numerical simulations is still not perfect, so it is now most common to simply use a fit to these results; fortunately, detailed simulations show that the resulting function can still *nearly* be phrased as a function of ν_c. Fits of the form

$$\frac{dn}{dM} = A' \sqrt{\frac{2a'}{\pi}} \frac{\rho_m}{M} \frac{-d(\ln\sigma)}{dM} \nu_c \left[1 + \frac{1}{(a'\nu_c^2)^{q'}}\right] e^{-a'\nu_c^2/2}, \tag{3.47}$$

which is closely motivated by ellipsoidal collapse, perform reasonably well. The best fit to simulations of this form is $a' = 0.75$ and $q' = 0.3$, and where proper normalization is ensured by adopting $A' = 0.322$. Refined numerical simulations

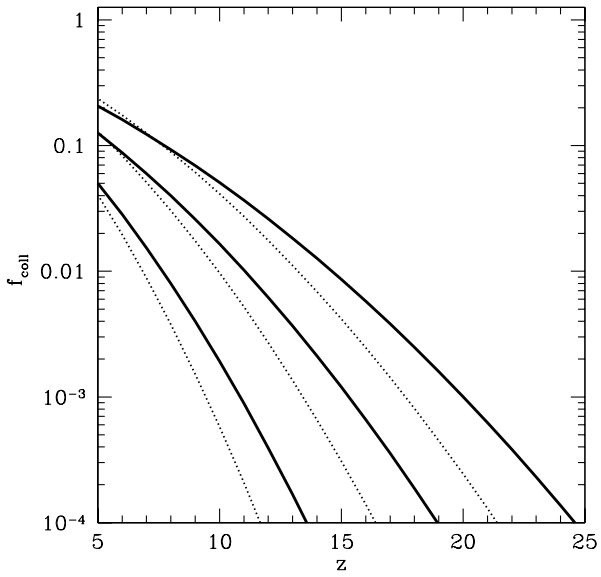

Figure 3.5 The collapse fraction of dark matter halos at high redshifts. The solid curves show $f_{\rm coll}$ computed from the mass function of eq. (3.47), which is motivated by ellipsoidal collapse with parameters determined by a fit to numerical simulations. From top to bottom the three curves show the fraction of mass in halos with $T_{\rm vir} > 10^3$, 10^4, and 10^5 K. The dotted curves use the original Press-Schechter form in eq. (3.39).

show that even more complex fits, in which the universal dependence on ν_c breaks slightly, are necessary, with the fitting parameters depending on redshift.[20] Results for the associated comoving density of halos of different masses at different redshifts are shown in Figure 3.4.

Figure 3.5 also shows $f_{\rm coll}$, or the fraction of mass above a given threshold (here shown as a function of $T_{\rm vir}$). The solid curves take the improved mass function in equation (3.47), while the dotted curves take the simpler (but less accurate) form of equation (3.39). The three sets of curves show the fraction in halos with $T_{\rm vir} > 10^3$, 10^4, and 10^5 K. We will see later that the middle value here corresponds to the threshold for efficient star formation before reionization, while the last is approximately the threshold for star formation after reionization. Note that, in all cases, $f_{\rm coll}$ increases extremely rapidly at high redshifts, since (at least at $z > 10$) all of these halos are far on the exponential tail of the mass function. Also note that the simple Press-Schechter prescription tends to underestimate the abundance of high-mass halos – and thus drastically underestimate $f_{\rm coll}$ when halos are rare – but slightly overestimates the abundance of low-mass halos (visible in the $T_{\rm vir} > 10^3$ K curves near $z \sim 5$).

3.6 HALO CLUSTERING IN LINEAR THEORY

To this point, we have computed the *average* abundance of halos throughout the universe. But of course the universe is not perfectly smooth on larger scales, and we naturally expect that large-scale overdensities have an overabundance of halos relative to the average, and that large-scale underdensities (or voids) have a deficit. This **clustering** is an extremely important aspect of halos in the real universe, especially at high redshift.

The excursion set formalism allows us to describe this clustering in detail, at least to linear order. A large-scale overdensity corresponds to $\delta > 0$ across a large but finite mass M_b. We then imagine our halos forming out of the material within this mass (similarly to a region carved out from a universe with a higher value of Ω_m). We can solve the related diffusion problem just as for the conditional mass function, simply by changing the origin of our random walks from $\delta = 0$ which is appropriate for $M \to \infty$ to δ within our region. The small head start these modified initial conditions provide to halos in overdense regions can be extremely important: recall that the density distribution itself is Gaussian, and so the abundance of rare halos is exponentially sensitive to the underlying density. We now wish to describe this dependence, often called the "peak-background split", quantitatively.

First, we should keep in mind that the Press-Schechter approach gives the *comoving* number density of halos or the number density of halos per unit mass rather than the more observationally relevant number per volume V. An overdense region with density $\delta_b = (\rho/\bar{\rho} - 1) > 0$ fills a smaller volume, by a factor $(1 - \delta_b)$. Thus, we expect the halo density to *appear* larger even if the total number of halos remained constant:

$$\left(\frac{\delta n}{n}\right)_{\rm halo} = \frac{V}{V(1-\delta_b)} - 1 = \delta_b. \tag{3.48}$$

This is the same factor by which the dark matter density itself changes, so if this were the only effect the halos would be an *unbiased* tracer of mass.

Next, we solve the usual diffusion problem with our modified initial conditions; for simplicity we will assume that M_b is sufficiently large to have $\sigma^2(M_b) \ll \sigma^2(M)$. As with the conditional mass function, the solution is identical to the usual form except that

$$\delta_{\rm crit} \to \delta_{\rm crit} - \delta_b \qquad \text{or} \qquad \nu_c \to (\delta_{\rm crit} - \delta_b)/\sigma(M). \tag{3.49}$$

We can therefore immediately write down the abundance in the region. However, it is most useful to consider a small overdensity $\delta_b \ll \delta_{\rm crit}$ and Taylor expand about the average result. Note that we have taken $\delta_{\rm crit}$ to be a function of redshift and used densities linearly extrapolated to the present day. We must therefore also extrapolate $\delta_b^0 = \delta_b(z)D(z)$. Expanding in a Taylor series:

$$\frac{dn}{dM}(\delta_b^{\rm ex}) \approx \frac{dn}{dM} + \frac{dn}{dM\,d\nu_c}\frac{d\nu_c}{d\delta_b^0}\delta_b^0 + ... \tag{3.50}$$

Using the original Press-Schechter mass function for simplicity, the halo abundance changes by a factor

$$\left.\frac{\delta n}{n}\right|_{\rm PS} = \frac{\nu_c^2 - 1}{\sigma\nu_c}\delta_b^0. \tag{3.51}$$

Canceling the growth factors that appear in both ν_c and δ_b^0, we obtain

$$\left.\frac{\delta n}{n}\right|_{\rm PS} = \frac{\nu^2 - 1}{\delta_c(z=0)}\delta_b(z). \tag{3.52}$$

Combining this effect with the change in volume (equation 3.48), we get

$$\frac{dn}{dM}(\delta_b) = \frac{dn}{dM}[1 + b(m)\delta_b], \tag{3.53}$$

where

$$b_{\rm PS}(M) = 1 + \frac{\nu_c^2 - 1}{\delta_c(z=0)}. \tag{3.54}$$

Obviously, because ν_c depends on mass implicitly through $\sigma(M)$, the bias also depends on the halo mass. Recalling that σ is a decreasing function of mass, we see that b will *increase* with halo mass: the abundance of larger halos fluctuates more than the abundance of small halos. This is because massive halos are on the exponential tail of the density distribution, so that the small boost provided by the overdense region has a large effect. Similarly, ν_c is an increasing function of redshift, so halos of a given mass become more biased earlier in cosmic history, when they are rarer. As a result, it is not simply the *total* abundance of halos that changes with background density: the *shape* of the mass function also changes, leaning more heavily toward massive halos in dense environments.

We have evaluated the bias for the Press-Schechter mass function; one can perform a similar exercise with the more accurate mass functions described in §3.5.3. For example, the mass function of equation (3.47) yields,

$$b_{\rm ST}(M) = 1 + \frac{q\nu_c^2 - 1}{\delta_c(z=0)}. \tag{3.55}$$

This result has the same qualitative trends as the earlier expression, although massive halos tend to be somewhat less clustered and small halos somewhat more. Figure 3.6 shows the bias for this model over the same mass and redshift ranges as Figure 3.4. Note that galaxy-mass halos ($M > 10^8 M_\odot$) can be quite highly biased during the era of the first galaxies, while very small halos are "anti-biased ($b_{\rm ST} < 1$) at late times. These halos tend to form in underdense regions, because those in overdense regions have already been swallowed by larger halos.

3.7 THE NONLINEAR POWER SPECTRA OF DARK MATTER AND GALAXIES

3.7.1 The Halo Model

We have now assembled several powerful ingredients in describing the distribution of matter in the Universe: *(i)* the mean abundance of halos as a function of mass and redshift; *(ii)* the clustering of these halos; and *(iii)* the radial structure of these halos (the NFW or Einasto profiles). The first two of these ingredients are constructed from linear theory; the third is taken from numerical simulations but is remarkably

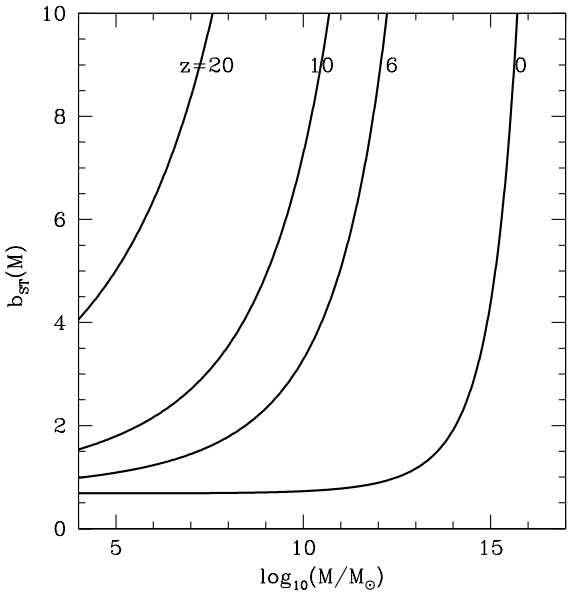

Figure 3.6 The linear bias of halos as a function of M at different redshifts z.

simple. We can now assemble these ingredients into a first stab at computing the statistical distribution of matter even in the *nonlinear* regime through a powerful approach known as the **halo model**.

The idea is to describe the power spectrum (or, alternatively, the correlation funciton) or dark matter by conceptually dividing all the matter in the Universe into halos of some – often very small – mass.[vi] Because the NFW profile describes the structure of *each* of these halos, and the excursion set formalism describes their abundance and statistical distribution, we can use this picture to compute the correlations between any two dark matter particles.

Before proceeding, we first write the NFW profile for a halo of virial mass M in the simplified form

$$\rho(r|m) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2}, \tag{3.56}$$

for $r < r_{\rm vir}$ and zero otherwise, where m labels the mass of the appropriate halo and $r_s = r_{\rm vir}/c_N$. We define a normalized density profile $u = \rho/M$, so that the integral over all space is unity. To compute the power spectrum we will work in

[vi] This follows naturally in the excursion set formalism, where any trajectory must *eventually* exceed an arbitrary threshold $\delta_{\rm crit}$ if $\sigma^2 \to \infty$ for $M \to 0$ in an arbitrarily-cold dark matter.

Fourier space; the Fourier transform of $u(r|m)$ is

$$u(k|m) = \frac{4\pi\rho_s r_s^3}{m}\left\{\sin(kr_s)\left[\mathrm{Si}([1+c]kr_s) - \mathrm{Si}(kr_s)\right] - \frac{\sin(kr_s)}{(1+c)kr_s} + \cos(kr_s)[\mathrm{Ci}([1+c]kr_s) - \mathrm{Ci}(kr_s)]\right\}, \quad (3.57)$$

where

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t} dt, \quad (3.58)$$

$$\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos t}{t} dt. \quad (3.59)$$

This is a rather unwieldy expression but can easily be computed numerically. To gain further insight, it is useful to consider halos with Gaussian density profiles and width r_s; then

$$u_\mathrm{g}(k|m) = \exp[-(kr_s)^2/2]. \quad (3.60)$$

This is near unity for $k \ll 1/r_s$ before falling off at larger wavenumbers; the shape of any realistic density profile is qualitatively similar.

Given our assumption that every dark matter particle lies within a halo, we can construct the total density field by simply adding up the profiles of all the halos:

$$\rho(\mathbf{x}) = \sum_i \rho(\mathbf{x} - \mathbf{x}_i|m_i) \quad (3.61)$$

$$= \sum_i m_i u(\mathbf{x} - \mathbf{x}_i|m_i) \quad (3.62)$$

$$= \sum_i \int dm \int d^3x' \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) m u(\mathbf{x} - \mathbf{x}'|m), \quad (3.63)$$

where i labels the different halos. In the last line, the integrals over mass and space simply fix the coordinates and mass of the halo in u.

Now note the useful identity

$$\left\langle \sum_i \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) \right\rangle = n(m), \quad (3.64)$$

where we write $dn/dm = n(m)$ for brevity. This happens because the Dirac delta functions in each volume appear a number of times equal to the number of halos (at each mass) per unit volume. So the mean density is

$$\bar{\rho} = \int dm \int d^3x' \left\langle \sum_i \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) \right\rangle m u(\mathbf{x} - \mathbf{x}'|m) \quad (3.65)$$

$$= \int dm\, m n(m). \quad (3.66)$$

3.7.2 The Correlation Function

Next, let us use the same approach to calculate the second moment, the correlation function, $\xi(\mathbf{x}-\mathbf{x}') = \langle\delta(\mathbf{x})\delta(\mathbf{x}')\rangle$. Here we have two integrals over space; the spatial average will act on

$$\left\langle \sum_{i,j} \delta(m_1 - m_i)\delta(\mathbf{x}_1 - \mathbf{x}_i)\delta(m_2 - m_j)\delta(\mathbf{x}_2 - \mathbf{x}_j) \right\rangle \tag{3.67}$$

Although analogous to the average in $\bar{\rho}$, this expression is much more complicated, and to evaluate it we must split it into two components, motivated by our picture of all particles sitting inside of halos. In that case, the two particles whose correlation we seek can have two qualitatively different configurations. First, they can sit inside the same halo (so that the labels are the same, $i = j$), in which case the halo density profile determines their correlation. Here we are summing twice over the same halo, so this part of the average is

$$n(m)\delta(m_1 - m_2)\delta(\mathbf{x}_1 - \mathbf{x}_2) \tag{3.68}$$

In the double integral, we therefore have

$$\begin{aligned}
&\int dm\, m^2 n(m) \int d^3x_1 \int d^3x_2\, \delta(\mathbf{x}_1 - \mathbf{x}_2) u(\mathbf{x} - \mathbf{x}_1|m) u(\mathbf{x}' - \mathbf{x}_2|m) \\
&= \int dm\, m^2 n(m) \int d^3x_1 u(\mathbf{x} - \mathbf{x}_1|m) u(\mathbf{x}' - \mathbf{x}_1|m) \\
&= \int dm\, m^2 n(m) \int d^3y u(\mathbf{y}|m) u(\mathbf{y} + \mathbf{x}' - \mathbf{x}|m) \equiv \bar{\rho}^2 \xi_{1h}(\mathbf{x} - \mathbf{x}')
\end{aligned} \tag{3.69}$$

where in the last line we let $\mathbf{y} = \mathbf{x} - \mathbf{x}_1$ and define the *one-halo correlation function* ξ_{1h}. We see that this term is the convolution of two density profiles, weighted by the halo's mass squared. This is just the integral over all pairs of particles within the halos.

The second possibility is that the two particles lie in separate halos; this case corresponds to the off-diagonal $i \neq j$ part of the double sum. The spatial average compares the locations of two halos of known separation, and it becomes

$$n(m_1)n(m_2)[1 + \xi_{hh}(\mathbf{x}_1 - \mathbf{x}_2|m_1, m_2)] \tag{3.70}$$

where ξ_{hh} measures the correlations between the halos themselves. Fortunately, we can easily compute this, at least when linear theory applies: we know the linear dark matter power spectrum and hence correlation function, $\xi(r)$, and we know from §3.6 how the halo densities reflect the underlying density. Therefore

$$\xi_{hh}(\mathbf{x} - \mathbf{x}'|m_1, m_2) = b(m_1)b(m_2)\xi_{\rm lin}(\mathbf{x} - \mathbf{x}'). \tag{3.71}$$

Note, however, that although equation (3.70) is general, equation (3.71) assumes that fluctuations in the halo distribution remain linear. This is not necessarily the case at high redshifts: even though the dark matter density fluctuations are very small, the halos can be so biased that the *halo* fluctuations are nonlinear (see Figure 3.6). One must be cautious with using linear theory in this regime.

In any case, these off-diagonal terms become

$$\int dm_1\, m_1 n(m_1) \int dm_2\, m_2 n(m_2) \int d^3x_1 \int d^3x_2\, u(\mathbf{x}-\mathbf{x}_1|m_1)u(\mathbf{x}'-\mathbf{x}_2|m_2) \times[1+\xi_{hh}(\mathbf{x}_1-\mathbf{x}_2|m_1,m_2)] \equiv \bar{\rho}^2+\xi_{2h}(\mathbf{x}-\mathbf{x}'), \quad (3.72)$$

where the *two-halo correlation function* ξ_{2h} describes correlations between particles in different halos. For some physics insight, suppose that halos are sharply peaked compared to the separation of interest, or $|\mathbf{x}-\mathbf{x}'| \gg r_{\rm vir}$. Then we can approximate the profiles as delta functions, and the integrals over x are easy. We therefore get

$$\xi_{2h}(\mathbf{x}-\mathbf{x}') \approx \xi(\mathbf{x}-\mathbf{x}') \int dm_1\, \frac{m_1}{\bar{\rho}} b(m_1) n(m_1) \int dm_2\, \frac{m_2}{\bar{\rho}} b(m_2) n(m_2) \quad (3.73)$$

This is just the normal dark matter correlation function ξ, weighted by the bias squared of all halos.

To compute this average, it is simplest to transform the integration variable to ν_c:

$$\int dm\, \frac{m}{\bar{\rho}} \left[1+\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] n(m) = 1+\int dm\, \frac{m}{\bar{\rho}} \left[\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] n(m) \quad (3.74)$$

$$= 1+\sqrt{\frac{2}{\pi}} \int d\nu_c \left[\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] e^{-\nu^2/2} \quad (3.75)$$

$$= 1, \quad (3.76)$$

where in second line we use $mn(m)dm = \bar{\rho}\sqrt{2/\pi}e^{-\nu^2/2}d\nu$. In hindsight, this is obvious: because all dark matter particles are in one halo or another, the net bias of the halo population relative to the dark matter must be zero!

Finally, combining the diagonal and off-diagonal terms we obtain the total nonlinear correlation function:

$$\xi(\mathbf{x}-\mathbf{x}') = \xi_{1h}(\mathbf{x}-\mathbf{x}') + \xi_{2h}(\mathbf{x}-\mathbf{x}'). \quad (3.77)$$

Again, this form has a simple physical interpretation: the net result is the sum of correlations of particles within halos, and those between halos. The relative importance of the two terms depends on the separation: when $|\mathbf{x}-\mathbf{x}'| \ll r_{\rm v}$, the particles sit inside a single halo so ξ_{1h} dominates; on much larger scales, ξ_{2h} is more important. On sufficiently large scales for linear theory to apply, the latter is very easy to compute in terms of the linear-theory dark matter correlation function.

3.7.3 The Power Spectrum

To obtain the power spectrum, we must simply take the Fourier transform of ξ. Because that is a linear operation, we again obtain separate one-halo and two-halo terms, with a total power spectrum

$$P(k) = P_{1h}(k) + P_{2h}(k). \quad (3.78)$$

The former is straightforward, since it is a simple convolution:

$$P_{1h}(k) = \int dm\, \frac{m^2 n(m)}{\bar{\rho}^2} |u(k|m)|^2. \quad (3.79)$$

The two-halo term is not as trivial. For simplicity, we will focus on the case in which equation (3.71) applies. First note that it is only a function of the separation, so we let $\mathbf{x} = 0$ and write (ignoring the integrals over mass and bias for now)

$$\xi_{2h}(\mathbf{r}) \propto \int d^3x_1 \int d^3x_2 u(-\mathbf{x}_1|m_1)u(\mathbf{r}-\mathbf{x}_2|m_2)\xi_{\rm lin}(\mathbf{r}) \tag{3.80}$$

$$= \int \frac{d^3k_1}{(2\pi)^3}u(\mathbf{k}_1|m_1)\int \frac{d^3k_2}{(2\pi)^3}u(\mathbf{k}_2|m_2)\int \frac{d^3k_3}{(2\pi)^3}P_{\rm lin}(\mathbf{k}_3)e^{i\mathbf{k}_2\cdot\mathbf{r}} \tag{3.81}$$

$$\times \int d^3x_1 e^{i\mathbf{x}_1(\mathbf{k}_1+\mathbf{k}_3)}\int d^3x_2 e^{-i\mathbf{x}_2(\mathbf{k}_2+\mathbf{k}_3)} \tag{3.82}$$

$$= \int \frac{d^3k_3}{(2\pi)^3}u(-\mathbf{k}_3|m_1)u(-\mathbf{k}_3|m_2)P_{\rm lin}(\mathbf{k}_3)e^{-i\mathbf{k}_3\cdot\mathbf{r}}. \tag{3.83}$$

In the second line, we took the Fourier transform of each piece and collected exponentials, and in the last we note that the integrals over $e^{i\mathbf{x}\cdot\mathbf{k}}$ are simply Dirac delta function. Inserting the mass integrals again, we have

$$P_{2h}(k) = P_{\rm lin}(k)\left[\int dm\,\frac{m}{\bar{\rho}}b(m)n(m)u(k|m)\right]^2, \tag{3.84}$$

where $P_{\rm lin}(k)$ is the linear theory power spectrum. Of course, when ξ_{hh} cannot be written simply according to equation (3.71), the expression for P_{2h} is more complex, although the general form of equation (3.78) still applies.

Let us summarize what we have accomplished. We began with the linear theory predictions for halo abundances and clustering. By simply adding the density profile of each halo (chosen from numerical simulations), the halo model ansatz allows us to compute the *nonlinear* power spectrum of dark matter from these linear theory predictions.

3.7.4 Nonlinear Bias

Figure 3.7 shows the resulting power spectra predictions at a range of redshifts as well as a comparison to the underlying $P_{\rm lin}(k)$. Not surprisingly, on sufficiently large scales the halo model prediction matches $P_{\rm lin}(k)$ precisely: on scales much larger than the halo size, $u(k|m) \to 1$. The factor in brackets in equation (3.84) then becomes the mass-averaged bias, which is just unity and so $P_{2h} \approx P_{1h}$. Meanwhile, on these scales $P_{1h} \approx$ constant, is small.[vii] At large k, the one-halo term – which describes the structures within each halo – dominates; it becomes more and more important as halos grow over time.

Unfortunately, Figure **??** also shows that, in comparison to detailed numerical simulations, the halo model prediction is not terribly accurate on the "crossover" scales between the one-halo and two-halo terms at high redshifts. This is because the assumption of linear bias breaks down in this regime. Although the dark matter fluctuations themselves are small at $z > 6$ (see Figure 2.1), massive halos are very highly biased (Figure 3.6). For example, at $z = 10$ the typical (linear theory)

[vii] In some applications of the galaxy power spectrum (see below), the constant value of P_{1h} may not be small when the objects of interest are not very rare.

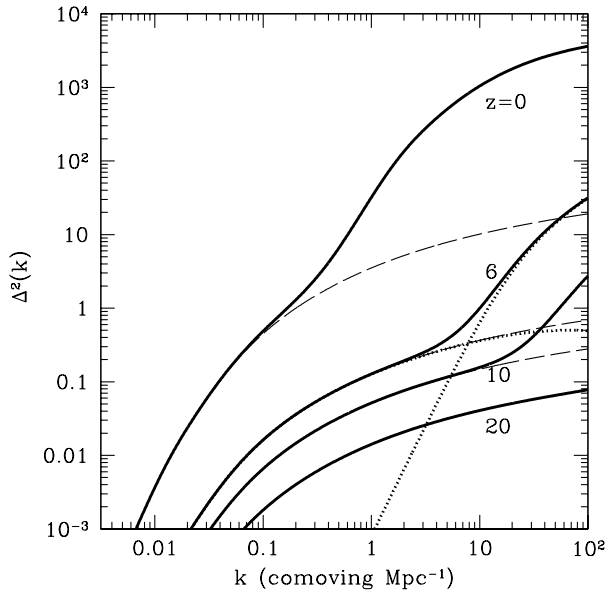

Figure 3.7 Dark matter density power spectrum predictions over a range of redshifts. The solid curves show the halo model prediction including only linear bias, while the dashed curves show the corresponding linear theory predictions. For the $z = 6$ curves, we also show separately the one-halo and two-halo terms with the dashed curves; these dominate at large and small k, respectively.

density fluctuation on a comoving scale of $2\pi/k \sim 1$ Mpc is $\sigma \sim 0.3$. But halos with $M > 10^8 M_\odot$ have $b_{\rm ST} > 3$, so their fluctuations are nonlinear. Although such halos contain only a small fraction of the mass, their nonlinear clustering is responsible for most of the structure at moderate to small scales.

One way to account for these nonlinear effects is to continue the expansion of equation (3.50) to higher order in δ_b. However, that is not easily incorporated into the halo model, because it requires incorporating higher order fluctuations – and the expansion becomes unwieldy once $b\sigma \sim 1$.

A more empirical approach is to allow for an "effective" scale-dependent bias $b_{\rm eff}(k|m)$, which is most often measured by comparison to numerical simulations. This function approaches the standard linear bias at small k but increases monotonically toward large k to reflect the nonlinear clustering of nearby halos.

One way to model this nonlinear bias is through the excursion set formalism. Recall that the correlation function (and hence power spectrum) represent the enhanced probability of finding a halo near another (usually because they lie in an overdense region). We can compute this probability within a *given* region of total mass M_b by assuming that the random walks for particles in the two putative halos in M–δ space are completely correlated on scales larger than M_b and uncorrelated on smaller scales. This gives the joint probability of finding two halos in the region, which provides a *scale-dependent* bias. The predictions of this model match $b_{\rm eff}$ from numerical simulations reasonably well.

3.7.5 The Galaxy Power Spectrum

The halo model approach is most often used to compute the power spectrum of galaxies (or, more specifically, subsets of galaxies that match an observable sample). Here we usually consider each galaxy to be a marker or signpost: we do not care whether the galaxy is large or small, just that it belongs to our statistical sample.[viii]

This necessitates just a few simple modifications to the formalism above. For example, the two-halo term (equation 3.84) has a factor of m inside each integral. This counted the number of dark matter particles inside each halo. Instead of counting pairs of particles, we only care about pairs of galaxies,

$$P_{2h}^{\rm gal}(k) = P_{\rm lin}(k) \left[\int dm \frac{\langle N|m\rangle}{\bar{n}_{\rm gal}} n(m) b_{\rm eff}(k|m) u_{\rm gal}(k|m) \right]^2 , \qquad (3.85)$$

where $\langle N|m\rangle$ is the mean number of galaxies in a dark matter halo of mass m and we have included the nonlinear bias correction $b_{\rm eff}$. We have also added two other small adjustments: we normalize to the average number density of galaxies in the sample, $\bar{n}_{\rm gal}$, and we include the profile of galaxies within the halo, $u_{\rm gal}$, rather than the dark matter density profile u.

Similarly, the one-halo term had a factor of m^2 reflecting the weighting of pairs

[viii] This is not a necessary condition of course: one could easily compute clustering statistics weighted by galaxy luminosity, for example. But that is rarely done in practice.

of particles within that halo,

$$P_{1h}^{\rm gal}(k) = \int dm\, n(m) \frac{\langle N(N-1)|m\rangle}{\bar{n}_{\rm gal}^2} |u_{\rm gal}(k|m)|^, \tag{3.86}$$

where $\langle N(N-1)|m\rangle$ counts pairs of galaxies.

Clearly, to compute the properties of a given sample we need an additional function which relates galaxies to dark matter halos. This *halo occupation distribution* can involve a great deal of the physics of galaxy formation, which we will discuss in later chapters. However, the basic principles are relatively simple; it is the application to real surveys that involves the subtleties. First, let us assume that each halo can have two types of galaxies: a "central" galaxy and satellites. The former typically exists if the halo exceeds some minimum mass threshold $M_{\rm min}$ (for example, the Jeans mass that we have already discussed, or the cooling mass that we will consider later); we can think of it as the halo's "initial" galaxy, tracing its history along the largest branch at each merger.

Satellites constitute the remaining population: they live inside "subhalos" that have not yet merged completely with the primary halo. Numerical simulations at low redshifts show that subhalos typically appear above some other minimum mass $M_{\rm sat} > M_{\rm min}$, and their number N_s increases roughly proportionally to the halo mass.

However, at high redshifts satellites are much less common: halos simply are not big enough to contain a substantial number of sub-galaxies, and halos are sufficiently small that even those with two merging galaxies may appear as a single irregular source in a real survey. Thus, for most purposes, we take $\langle N|m\rangle = 1$ if $m > M_{\rm min}$ and zero otherwise.

It is somewhat more difficult to compute the one-halo term. At low redshifts, the number of satellites is found to be roughly a Poisson variable (which is reasonable, since merging is a somewhat stochastic process), so that $\langle N_s(N_s-1)\rangle = \langle N_s\rangle^2$. This implies that $\langle N(N-1)\rangle = \langle N\rangle^2 - 1$, which is sub-Poisson at the low-mass end. In the high-redshift limit, where satellites are unimportant, the one-halo term disappears because there is only one galaxy.

In addition to these terms, which arise because galaxies trace the density field, we must also add in stochastic "shot-noise" fluctuations arising from the discrete nature of galaxies: any such measurement is fundamentally a counting exercise, so we expect Poisson errors in the galaxy number counts to provide an additional source of fluctuations. In a volume V, the variance in the galaxy number counts is therefore $\sim nV$, so the fractional density fluctuation in a mode with wavenumber k will be $\Delta^2_{\rm shot} \sim 1/nV \sim k^3/n$. A more precise derivation shows $P_{\rm shot}(k) = 1/n$,[21] or

$$\Delta^2_{\rm shot} = \frac{k^3}{2\pi^2 n}. \tag{3.87}$$

This noise term contains no interesting physics and must be removed from an observed power spectrum in order to study the interesting physical component tracing the underlying density field. Fortunately, that is usually easy, so long as one has a reasonable estimate for the sample's true number density (i.e., $nV_{\rm survey} \gg 1$).

3.8 NUMERICAL SIMULATIONS OF STRUCTURE FORMATION

Although the models we have discussed in this chapter are useful, they inevitably fall short of a complete description of the structure and dynamics of dark matter and baryons in an expanding Universe. Each dark matter particle responds to the gravitational force from every other particle within its causal horizon, and the dynamics of baryons is also affected by their gas pressure gradient (and interaction with photons). A comprehensive description of this problem is far beyond the capabilities of any analytic model.

Fortunately, the rapid increase in computing power over the past several decades has enabled numerical calculations to address this challenge. Computers are particularly well-suited to this endeavor, because they can easily calculate the simple physical interactions between many particles. Although following the behavior of individual dark matter particles is still not feasible, numerical simulations can now (as of 2011) follow the dynamics of $\sim 10^{10}$ particles over long periods of cosmic history. The fundamental idea behind cosmological numerical simulations is to discretize the density field $\rho(\mathbf{x})$ into a large number of particles or grid cells and follow their evolution, incorporating as many physical processes (preferably from first principles) as possible. This allows detailed comparisons of theoretical predictions with observations as well as the study of "emergent phenomena" that depend upon the interaction of many physical inputs and so cannot easily be predicted from analytic models. Nevertheless, one must always bear in mind that a numerical simulation is ultimately no better than the physics underlying its component algorithms, and it is crucial to understand those inputs in order to assess the the reliability of linking simulation results to observables in the sky.

Numerical simulations have been instrumental to understanding large scale structure, the Lyman-α forest, the formation of the first stars, and a number of other topics that we will discuss. In the remainder of this section, we will briefly discuss their most important features and limitations. We will, however, defer discussion of computational radiative transfer to chapter **??** and focus here on gravitational and gas dynamics.

3.8.1 Gravitational Dynamics: N-Body Codes

The simplest problem is to follow the gravitational interactions of cold collisionless particles in an expanding Universe. If we have a collection of N particles with particle mass m, each labeled by index i and a comoving position and peculiar velocity $(\mathbf{x}_i,\, \mathbf{u}_i)$, this amounts to solving the equations of motion (c.f., equations 2.2-2.3)

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{u}_i \tag{3.88}$$

$$\frac{d\mathbf{u}_i}{dt} + 2H(t)\mathbf{u}_i = -a^{-2}\nabla\phi, \tag{3.89}$$

where the gravitational potential is determined by the Poisson equation (2.4). To solve this problem, we discretize time into a sequence t_n and assume that we know the initial values $[\mathbf{x}_i(t_1),\, \mathbf{u}_i(t_1)]$ for all particles. Then the future configuration can be solved by numerically integrating the above system.

The key point is to determine the force on each particle; the crucial element involves choosing an integration scheme that is both stable and resistant to secular numerical errors. The simplest such scheme is known as a *leapfrog* approach, because it uses two different sets of discretized times for the different input quantities. For example, suppose we know the position at a time t_n and wish to know it after a single timestep, at $t_{n+1} = t_n + \Delta t$. As an intermediate step, we compute the position and acceleration $\mathbf{a}_i$ of particle i at $t_{n+1/2} = t_n + \Delta t/2$:

$$\mathbf{x}_i(t_n + \Delta t/2) = \mathbf{x}_i(t_n) + \mathbf{u}_i(t_n)\Delta t/2, \tag{3.90}$$

$$\mathbf{a}_i(t_n + \Delta t/2) = \mathbf{a}[\mathbf{x}_i(t_n + \Delta t/2), t_n + \Delta t/2], \tag{3.91}$$

where the acceleration at the intermediate time depends upon the predicted location of the particle then as well as the locations of all the other particles. We can then compute the new position and velocity at the final time t_{n+1} using the acceleration at the intermediate time,

$$\mathbf{u}_i(t_{n+1}) = \mathbf{u}_i(t_n) + \mathbf{a}_i(t_n + \Delta t/2)\Delta t, \tag{3.92}$$

$$\mathbf{x}_i(t_{n+1}) = \mathbf{x}_i(t_n) + [\mathbf{u}_i(t_n) + \mathbf{u}_i(t_{n+1})]\Delta t/2. \tag{3.93}$$

This is superior to Eulerian integration schemes because, by evaluating the acceleration at the midpoint of the timestep, it improves time-reversibility and better preserves the phase space properties of the particle orbits.

This scheme requires computing

$$-\nabla\phi(\mathbf{x}_i) = -Gm\sum_{j\neq i}\frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|^3}. \tag{3.94}$$

for each particle. However, the computational time required to calculate all these forces scales as N^2 and is prohibitive even for modest size systems. Modern codes use one or more tricks to simplify the calculation. The most straightforward is a **tree algorithm**, which groups distant particles into sets (with the group size generally increasing for more distant particles). The gravitational force from each group can then be estimated using a multipole expansion. Grouping algorithms can speed up the calculation to scale with particle number as $N \log N$.

A second trick is to use a Fast Fourier Transform (FFT) algorithm to compute the force on a grid, a technique known as a **particle-mesh (PM) algorithm**. In this approach, the particle mass distribution is smoothed and mapped onto a uniform mesh. Poisson's equation can then be solved rapidly via an FFT, provided that the computational box is assumed to have periodic boundary conditions. The force at each grid point follows via an inverse Fourier transform. Finally, the force at each particle location is computed via interpolation. This approach scales linearly with particle number, but the practical limit is often dictated by computer memory, since the mesh resolution ultimately determines the force accuracy. This approach does not, however, deal with highly clustered particles very well. A compromise, called $\mathbf{P^3M}$, adopts direct summation at small separations and a series of adaptive grids, is often used instead.

There are three effective resolution limits on N-body codes. The first is the particle mass m, which obviously determines the smallest object that can be resolved.

Typically, $> 10^3$ particles are required to determine the density profile of a virialized halo reliably, and many orders of magnitude more to resolve its substructure in detail.

A second limit emerges from the discretization of the density field: the point mass force calculation in equation (3.94) causes large artifical deflections when particles pass very close to each other. This is unphysical because the particles should actually be distributed over larger volumes. To alleviate this problem, codes introduce a *force-softening* parameter such that the force scales as $1(r^2 + \varepsilon^2)$. This limits the maximum resolvable density contrast to $(\bar{\ell}/\epsilon)^3$, where $\bar{\ell}$ is the mean particle spacing. For most modern codes, this lies in the range $\sim 20^3$–50^3. For PM codes, the force limit is roughly twice the grid spacing, because it depends on the gradient of the potential across that grid.

The final limit comes form the requirement that the numerical integration time remain stable. Crudely, this requires that the timesteps be sufficiently close so that the first-order approximations intrinsic to the integration converge. Equivalently, the series

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \mathbf{u}(t_i)\Delta t + \frac{1}{2}\mathbf{a}(t_i + \Delta t/2)\Delta t^2 + ... \tag{3.95}$$

must converge rapidly. Here the force per unit mass is typically evaluated at the midpoint of the particle's trajectory in order to improve stability and convergence. The ratio of these terms suggests

$$\Delta t = k\frac{\sigma}{|\mathbf{a}|}, \tag{3.96}$$

where $k < 1$ is an imposed tolerance parameter and σ is the typical velocity dispersion of the particles in the simulation. This is known as the *Courant-Friedrichs-Lewy condition* (often referred to as the Courant condition).

3.8.2 Hydrodynamics: Grid-Based Approaches

Extending the calculation beyond dark matter increases significantly its complexity, because the trajectories of baryons are shaped by hydrodynamic forces in addition to gravity. In particular, converging gas flows can lead to the development of sharp discontinuities (shock fronts) whose accurate treatment requires high spatial resolution. The complete fluid equations can be written (in proper coordinates) as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{3.97}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla\Phi - \frac{1}{\rho}\nabla p, \tag{3.98}$$

$$\frac{\partial \varepsilon}{\partial t} + \mathbf{v} \cdot \nabla\varepsilon = -\frac{p}{\rho}\nabla \cdot \mathbf{v} + \frac{(H - Q)}{\rho}. \tag{3.99}$$

These represent the conservation of mass, momentum, and specific energy (per unit mass) ε, respectively. In the last equation, H and Q are the radiative heating and cooling rates per unit volume, respectively, and we have ignored any other internal

heating mechanism. Alternatively, the energy equation can be replaced with an equation for the entropy per unit mass s,

$$\rho T \left(\frac{\partial s}{\partial t} + \mathbf{v} \cdot \nabla s \right) = (H - Q). \tag{3.100}$$

We have written these equations in an *Eulerian* form, in which the spatial coordinate system is fixed. An alternative is a *Lagrangian* approach, in which the coordinates move with the fluid elements. In this case, the appropriate derivative is the **convective derivative**,

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla, \tag{3.101}$$

so that, for example, equation (3.98) simplifies to

$$\frac{D\mathbf{v}}{Dt} = -\nabla\Phi - \frac{1}{\rho}\nabla p. \tag{3.102}$$

There are two common approaches to solving this system of equations. The first is to divide space into a uniform grid and to solve the hydrodynamic equations for cell-averaged quantities at each grid point. This Eulerian scheme is attractive because the mass, momentum, and energy components of the fluid equations can all be cast as flux conservation laws

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{F} = 0, \tag{3.103}$$

where q is the (cell-averaged) density ρ, momentum density $\rho u_{x,y,z}$, or total energy density $\rho(\varepsilon + u^2/2)$ and $\mathbf{F}$ represents the flux of this conserved density across the cell boundaries. This formulation lends itself naturally to grid-based methods: it means that to track the evolution of q at a particular location we need only keep track of the flux $\mathbf{F}$ through each of the cell boundaries. Labeling cells by an index k and assuming they are $\Delta\ell$ across, we have

$$q_k(t + \Delta t) = q_k(t) + \frac{\Delta t}{\Delta x} \sum_{\ell=1}^{3} [F_{\ell+}(t) - F_{\ell-}(t)], \tag{3.104}$$

where the indices ℓ label the three axes of the cells and $F_{\ell+}$ represents the flux along the i direction between the cell of interest and the next cell in the positive ℓth direction (and $F_{\ell-}$ the flux in the negative ℓth direction). Formulating the fluid equations in this way has one important advantage over the usual differential forms mentioned earlier: when fluids develop sharp discontinuities, like shocks, the latter break down. Instead, the fluid must be followed with integral forms like equation (3.103), which in the case of shocks are known as the Rankine-Hugoniot jump conditions.

The subtlety in grid-based methods lay in ensuring numerical stability for the solutions. For example, the most naive approach to estimate a fluid variable Q at a cell interface is to simply take an average with the cell-averaged quantities in the neighboring cell (e.g., $Q_{k+} = [Q_{k+1} + Q_k]/2$). However, this simple approach is in fact unstable, and more sophisticated algorithms are required. One common strategy is to approximate the calculation as a *Riemann problem* (also known as

a shock tube), in which a fluid quantity is constant over two regions with a discontinuity in between them. Provided that the system obeys conservation laws of the form in equation (3.103), Riemann problems can be solved exactly in terms of *characteristics* that propagate at known speeds in either direction; this exact solution can then be leveraged to efficiently calculate the evolution in more realistic circumstances. (For example, an initially uniform gas with a sharp edge adjacent to vacuum would flows into the vacuum at the sound speed, while a rarefaction wave would travel in the opposite direction through the gas, also at the sound speed.)

One popular technique for leveraging the Riemann problem is known as Godunov's scheme. One approximates each cell by its average value and then solves the Riemann problem at each of its interfaces. The resulting waves can then be propagated into the cell and its new properties calculated at a later time. In order to avoid the waves colliding and interacting with each other, the timestep must be limited by

$$\Delta t = k_{\rm grid} \frac{(\Delta x/2)}{c_s}, \tag{3.105}$$

where c_s is the sound speed and $k_{\rm grid}$ is a dimensionless constant.

Another example is the *piecewise parabolic method (PPM)*, which uses a parabolic function to interpolate a fluid variable across a cell and its immediate neighbors (it is thus a third-order extension of the basic Godunov method). The algorithm is constructed so as to mimic the propagation of nonlinear waves in the fluid system and to accurately capture shocks. Unfortunately, interpolation can also induce spurious oscillations when the fluid quantities change rapidly (as they do, for example, in shocks). These too can make the solutions unstable. One can therefore introduce a numerical dissipation scheme to damp these fluctuations, or alternatively enforce a *flux (or slope) limiter* that forces spatial derivatives to remain within reasonable bounds.

The disadvantage of grid-based approaches is that the grid resolution must be uniform, whereas the *desired* resolution may vary across the simulation volume – for example, the relevant spatial scales are much smaller near a collapsed dark matter halo than in a large void. Thus, one "wastes" computational resources in some regions. A common solution to this problem is **adaptive mesh refinement (AMR)**, in which finer grids are introduced as necessary to sub-volumes of the computation.

The fundamental idea of AMR is to demand that the local grid spacing adjust "on-the-fly" to the physical conditions within the fluid. For example, if a dark matter halo collapses to high density and accretes baryons, the physical resolution must increase in order to follow the flow. Meanwhile, the timestep required with a smaller grid will shrink dramatically according to equation (3.105). AMR codes spawn smaller meshes that are stepped at higher rates, while the background grid continues its slow evolution in low-density regions. While AMR does allow a dramatic increase in the dynamic range of grid-based calculations, the spawning of grids is an imperfect process that leads to some subtle numerical problems when the resolution increases discontinuously, especially in populating the initial conditions of small-scale modes originally absent from the parent grid.

While AMR solves the most glaring problem with grid-based approaches in astrophysics, these codes suffer from some other important shortcomings. Foremost amongst them is the lack of Galilean invariance inherent to such methods: because the advection terms in equations (3.97)–(3.99) are modeled explicitly, they inevitably contain numerical errors that depend upon the magnitude of the bulk velocity relative to the velocity dispersion (which can be large – for example, galaxies merge at velocities comparable to or greater than their own velocity dispersions). This creates numerical viscosity and diffusion that violate Galilean invariance. Without large physical transport coefficients, these numerical artifacts are in fact the leading order terms, so even the qualitative solutions may be questionable under some circumstances. In general some amount of dissipation is helpful, but limited resolution or high bulk velocities will cause over-mixing. Similar artifacts also appear whenever the bulk velocity is much larger than the thermal velocities; these can be remedied with a careful choice of the reference frame but not entirely removed.

3.8.3 Hydrodynamics: Particle-Based Methods

The alternative to grid-based approaches, **smoothed particle hydrodynamics (SPH)**, discretizes the fluid field and implicitly adapts the resolution to the local fluid properties. It is more naturally suited to problems with a high dynamic range of density, but it also faces its own set of challenges.

SPH methods formally aim to recover a smoothed version Q_s of a fluid field Q,

$$Q_s(\mathbf{r}) \equiv \int d^3\mathbf{r}'\, Q(\mathbf{r}) W(\mathbf{r}-\mathbf{r}', h), \tag{3.106}$$

where $W(\mathbf{r}, h)$ is a smoothing kernel, with h describing its characteristic width. Most commonly, this kernel has a cubic spline form with $W(\mathbf{r}, h) = w(r/2h)$ and

$$w(x) = \frac{8}{\pi} \begin{cases} 1 - 6x^2 + 6x^3, & 0 \le x \le 1/2, \\ 2(1-x)^3, & 1/2 \le x \le 1, \\ 0, & 1 < x. \end{cases} \tag{3.107}$$

Note that each particle therefore has a finite "width" $2h$ in this scheme.

Now suppose that we know the fluid properties at a set of points $\mathbf{r}_i$. We associate particles with each of these points, assigning mass m_i so as to conserve the total mass in the field and densities ρ_i such that the volume between the particles is $\sim m_i/\rho_i$. We can then estimate the smoothed field Q_s by a summation over these particles, so that equation (3.106) becomes

$$Q_s(\mathbf{r}) \approx \sum_k \frac{m_k}{\rho_k} Q(\mathbf{r}_k) W(\mathbf{r}-\mathbf{r}_k, h). \tag{3.108}$$

This sum is accurate so long as the kernel width h exceeds the (local) particle spacing. More precisely, one can set the density of particle i as[ix]

$$\rho_i = \sum_{k=1}^{N_{\rm ngb}} m_k W(\mathbf{r}_i - \mathbf{r}_k, h_i), \tag{3.109}$$

[ix] Note that we choose one scheme here for concreteness, but others are sometimes used as well.

where h_i is set so as to ensure that each particle has a fixed "mass" $\rho_i h_i^3$ =constant. This ensures that the number of neighbors $N_{\rm ngb}$ within its kernel is also nearly constant. This is the key advantage of SPH approaches: it can automatically adjust the degree of smoothing to the density of particles, focusing the "high-resolution" part of the calculation in volumes where it is most needed.

Equation (3.108) is generally taken as the SPH estimate for any fluid field. The derivatives of such a field can then easily be calculated (as they require only the derivatives of the kernel W), and from them one can construct discretized versions of the fluid equations. For example, equation (3.97) becomes

$$\frac{D\mathbf{v}_i}{Dt} = -\nabla\phi - \sum_{k=1}^{N_{\rm ngb}} m_k \left(\frac{p_i}{\rho_i^2} + \frac{p_k}{\rho_k^2}\right) \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h), \tag{3.110}$$

where ∇_i is the gradient with respect to the $\mathbf{r}_i$ coordinates. Unfortunately, this straightforward approach contains a number of subtleties in its practical application regarding bookkeeping between particles, smoothing lengths, etc. Here that is reflected in the loose notation $W(\mathbf{r}_i - \mathbf{r}_k, h)$, which does not specify the smoothing length to be used in the derivative (namely whether it applies to particle i or k).

The most popular astrophysical codes therefore take a slightly different approach by noting that the fluid equations (3.97)–(3.99) follow from the Lagrangian

$$\mathcal{L} = \int d^3\mathbf{r}\, \rho \left(\frac{|\mathbf{v}|^2}{2} - \varepsilon\right), \tag{3.111}$$

which itself can be easily discretized,

$$\mathcal{L}_{\rm SPH} = \sum_i \left(\frac{m_i |\mathbf{v}_i|^2}{2} - m_i \epsilon_i\right), \tag{3.112}$$

where the thermal energy of a given particle is assumed to depend only on its entropy. For now we will assume that entropy to be constant (i.e., we will neglect shocks and other dissipative processes).

The advantage of this Lagrangian formulation is that it can straightforwardly incorporate the constraint $\rho_i h_i^3$ =constant to define the smoothing length, as in any system of particles in elementary mechanics. Following the standard Lagrangian procedure, the equation of motion for this system is

$$\frac{D\mathbf{v}_i}{Dt} = -\nabla\phi - \sum_{k=1}^{N_{\rm ngb}} m_k \left(f_i \frac{p_i}{\rho_i^2} \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h_i) + f_k \frac{p_k}{\rho_k^2} \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h_k)\right), \tag{3.113}$$

where

$$f_i \equiv \left(1 + \frac{h_i}{3\rho_i} \frac{\partial \rho_i}{\partial h_i}\right)^{-1} \tag{3.114}$$

arise from the constraint. Note the similarity to equation (3.110); this slightly more complicated form implicitly includes the particle accounting without much increased complexity, and the direct derivation from a discretized Lagrangian manifestly conserves linear momentum, angular momentum, and energy.

Although elegant, this approach has one key weakness: namely, by assuming a constant entropy it does not allow shocks or other forms of dissipation. The above equation must then be supplemented with an *artificial* viscosity that re-introduces these features. Perhaps surprisingly, it is relatively easy to formulate this viscosity in such a way that it generates the proper additional entropy at shocks, so long as the prescribed viscosity conserves momentum and energy. This follows because the shock jump conditions (and hence macroscopic fluid variables) are independent of the transport coefficients such as the viscosity. However, SPH codes cannot resolve the structure of the shock itself unless the viscosity parameter reflects the microphysics of the gas; typically, shocks are much broader in SPH treatments than in grid-based codes. Another challenge is to ensure that this artificial viscosity does not affect the dynamics in regions outside of shocks.

The SPH approach requires a time integrator; because the fluid has been discretized into particles, the same leapfrog methods described in §3.8.1 work equally well here (though note that the irreversibility of most hydrodynamics process actually means that other methods work as well). The time steps must respect the Courant condition of equation (3.96), but because the hydrodynamics equations also involve spatial derivatives an additional limit applies as well, with $|v\Delta t|/\Delta r < 1$. This is usually written as

$$\Delta t_i = k_{\rm SPH}\frac{h_i}{c_{s,i}}, \tag{3.115}$$

where $c_{s,i}$ is the sound speed at the location of the ith particle. In practice, because the particle sizes and sound speeds can vary dramatically in a cosmological system, most codes allow for different particles to have different timesteps. This too allows the calculation to spend the bulk of its resources where they are most needed.

In addition to the difficulties with resolving shocks, SPH codes also have some problems following certain important fluid instabilities, like the Kelvin-Helmholtz instability in shear flows. Particle-based schemes inevitably contain "noise" in their realizations of the density and velocity fields, which in certain regimes can cause unphysical effects such as preventing the Kelvin-Helmholtz instability from growing. The noise can be suppressed by introducing an artificial viscosity that smooths the fluid fields, but that viscosity itself affects the instabilities as well. Clearly, one must pay careful attention to matching the ideal computational method for any particular physical problem.

It is also worth noting that, although SPH simulations do intrinsically adapt to high-density environments, they cannot "zoom" indefinitely. Once the timestep of equation (3.115) is too short – say in runaway gravitational collapse – it becomes impractical to continue the integration. The problem can be circumventing by creating a *sink particle* that accretes mass (and possibly exerts feedback in some prescribed manner) but whose internal structure is not resolved. This technique is used in simulations of star formation and is an example of *subgrid models* that represent physical processes unresolved by the simulation itself (see §3.8.4 for more discussion of these approaches).

Finally, SPH is ill-suited to problems in which the mixing of different fluids is important (such as diffusion), because the particles are generally not allowed to

exchange mass. This has more important ramifications than simply following mass around, however: entropy generation through gas mixing is impossible to follow reliably with standard SPH codes.

Although SPH is by far the most popular particle-based solver, it is not the only approach; the kernel is fundamentally used only to partition the fluid field into mass elements, and other schemes to accomplish the same purpose can also be used. For example, one can compute a *Voronoi tesselation* for the volume. This assigns a volume to each particle that includes all regions closer to its location than to any other particle, without any overlap between the particles. The same Lagrangian technique described above works with this modified constraint to write the equations of motion for each particle.

A step beyond that is to combine the advantages of particle and grid-based approaches by constructing a "moving mesh" of grid cells using the Voronoi tessellation technique to build the cells. Because the cell boundaries are well-defined (unlike in SPH), grid-based numerical algorithms can be used to compute the fluxes of integral fluid properties across the Voronoi mesh cells. Codes exploiting these techniques (such as AREPO) are just now becoming available.

3.8.4 The Limits of Numerical Simulations

Computational astrophysics has risen dramatically in importance over the last several decades, and the continuing increase in computing power promises to make these methods even more useful in the future. They have been instrumental in shaping our understanding of many aspects of astrophysics, including the high-redshift Universe. Nevertheless, one should keep in mind that they represent one tool in our arsenal for addressing challenging problems, and they rarely provide a complete physical understanding of such problems. It is therefore important to identify their limitations for any particular set up and to calibrate the significance of their results in that context.

We have already discussed some of the specific computational challenges that the different codes face: for example, grid-basd codes typically violate Galilean invariance and have difficulty with supersonic flows, while SPH codes do not resolve shocks properly and do not follow shear instabilities well. We have also discussed how the finite grid size or particle number limits the spatial resolution that any particular simulation can probe (though in a predictable manner). But astrophysical applications present deeper problems as well.

Foremost amongst them is the enormous dynamic range required to simulate cosmological volumes from "first principles." Ideally, we would like a simulation that resolves star formation inside dwarf galaxies but also samples a representative volume of the intergalactic structures. We will see in chapter 8 that, during cosmological reionization, this requires sampling a volume $> (100 \text{ Mpc})^3$. Meanwhile, star formation occurs down a scale $\sim R_\odot = 2.3 \times 10^{-14}$ Mpc. Covering both at once requires a spatial dynamic range $\sim 10^{16}$, far beyond the capabilities of even the largest computer clusters today or in the foreseeable future.

Cosmological simulations must therefore inevitably incorporate **subgrid models** to approximate physics unresolved by the simulation. The importance of these

prescriptions depends on the dynamic range and goals of the simulation. Most commonly, they parameterize processes inside galaxies, including:

- *Star formation:* Cosmological simulations, and even simulations of individual galaxies, are far from being able to resolve star formation – and, as discussed in later chapters, we are still far from understanding that process even after zooming to very small scales. Simulations must therefore construct a subgrid model for star formation, usually calibrating it to an empirical relation such as the Kennicutt-Schmidt law (see §9.5.4). The computed star formation rates are therefore no more reliable than the empirical or semi-analytic model underlying the simulation.

- *Black hole growth:* An equally difficult problem is the accretion of gas onto black holes, which typically occurs on solar system scales inside the complex environments of galactic nuclei. Without resolving the detailed gas dynamics at the center of galaxies (which is possible in specialized simulations, but not in their cosmological scale counterparts), it is impossible to determine the accretion rates onto these objects from first principles. It is therefore necessary to impose a subgrid model in order to track the growth of black holes and quasar activity.

- *Galactic winds and feedback:* We will see in chapter 6 (and §9.5.6) that feedback is likely ubiquitous in star-forming galaxies and crucial for regulating their star formation rates. The energy and momentum injected from supernovae and radiation likely prevents much of the gas from cooling into stars and removes material from the galaxy, likely enriching the intergalactic medium (IGM) with metals. However, these processes are difficult to model even in very high-resolution simulations, and simple prescriptions are usually implemented in cosmological simulations. The free parameters are then calibrated to local observations of feedback on galactic scales.

  Even more difficult to model is feedback from supermassive black holes, which can be very important energetically but has very limited observational constraints. Because it occurs most often at the centers of galaxies, the transport of the energy and momentum through the galaxy is crucial for model it effectively. For example, nearby radio galaxies launch powerful jets into the IGM, but it is not clear that these jets couple strongly to their host galaxies. With only a crude physical understanding of these processes, subgrid models that make strong assumptions about the underlying coupling mechanisms (in the form of relativistic and non-relativistic outflows, radiative heating, radiation pressure, or cosmic rays) are necessary.

- *Clumping:* We will see in chapter 8 that small-scale gas clumping is crucial to understanding reionization, but many cosmological simulations do not resolve the relevant physical scales (especially before reionization, when the Jeans mass is small). Moreover, this small-scale structure will evolve as the IGM temperature and pressure change. Often a subgrid model is inserted

to describe this clumping: it can include the clumping from unresolved filaments and sheets in the cosmic web (see chapter **??**) as well as the photoevaporation of collapse "minihalos" that are unable to form stars because their low virial temperature does not allow the gas to cool further. Some reionization simulations ignore hydrodynamics entirely and impose *all* gas clumping through a simplified prescription.

- *Radiative heating and cooling:* For most of the baryons in the universe, radiative processes – either photoheating from ionization or cooling from line transitions – are amongst the most important mechanisms setting their thermal properties. These in turn depend not only upon the metagalactic radiation field (which must be imposed externally, unless radiative transfer is included), but also on unresolved physics of the gas, including its metal content and any multiphase medium. Although coarse resolution likely suffices in the IGM, gas near or inside galaxies is subject to major uncertainties from these effects.

The importance of these subgrid models cannot be understated: nearly all of the *observable* predictions of cosmological simulations rely on their parameterizations. Indeed, it is no coincidence that the most influential cosmological simulations have often not been those with the most computing power; instead, they have made the most important advances in implementing physically motivated subgrid models.

Another problem, particularly at high redshifts is ensuring that the simulation samples a representative volume of the Universe. Typically, this is ensured by demanding that the largest density modes in the simulation remain well in the linear regime at the time the simulation is ended. For technical reasons (in order to make a Fast Fourier Transform easy, and so that the density field "outside" of the box can be represented by the box itself), most cosmological simulations implement *periodic boundary conditions*, in which opposite faces of the box are identified with each other. This forces the mean density of the box to take on the average cosmological value, which at first blush automatically appears to make the box a "representative" volume of the Universe. However, for highly clustered objects (which includes galaxies at very high redshifts), this may be misleading, because even a small density boost in a long-wavelength mode can dramatically affect the halo abundance. For sufficiently rare objects, most such objects may actually lie inside large-scale overdensities; a periodic box at the mean density can therefore not contain a fair sample of these halos. Fortunately, this effect can easily be quantified using the conditional mass function in the excursion set formalism.

For similar reasons, rare objects (like extremely massive halos) are very difficult to simulate, although they are the most interesting because their extreme properties often make them the easiest to observe. Typically, one studies such an object with an adaptive technique, although SPH and AMR on their own are rarely up to the task. Instead, an object of interest is identified (but not resolved sufficiently) in a large-scale simulation, and then another higher-resolution simulation is performed using the object's large-scale environment as boundary conditions.

Computer simulations are no more intelligent than their creators, and they rely on the proper input physics in order to produce reliable answers. Their construction

and proper use therefore requires as broad and deep a physical understanding as any other area of theoretical astrophysics. Computers follow the algorithms with which they are programmed, and they are limited by the approximate sub-grid physics that was implemented into them. They are therefore most effective at identifying and understanding so-called *emergent phenomena*, in which complex systems grow from the interactions of simple systems whose physics can individually be accurately described, or in making high-precision predictions for well-understood phenomena. However, if the input physics is incorrect – if the code uses incorrect initial conditions, or excludes any important physical process – the simulation is no better than an analytic model with similar flaws. A recent example is the recognition of baryonic acoustic oscillations – nominally a second order effect and so ignored in cosmological simulations of structure formation – as potentially providing a crucial modulation of the collapsed matter field (see §3.3).

In many astrophysical problems, these inputs are so poorly understood that a computer simulation is no better than a simple toy model (and, most likely, much less flexible). We urge the reader to combat the natural human tendency to conflate accuracy with precision: a computer is capable of blindly following incorrect physical assumptions toward an incorrect – but highly precise – solution (often accompanied by beautiful pictures and animations). It is important for both observers and theorists to appreciate the strengths and limitations of any theoretical calculation in detail before comparing its predictions to other calculations or observations.

Chapter Four

The Intergalactic Medium

4.1 THE COSMIC WEB

Although much of astronomy focuses on the luminous material inside galaxies, the majority of matter today – and the vast majority at $z > 6$ – actually lies outside of these structures, in the *intergalactic medium* (IGM). This material ultimately provides the fuel for galaxy and cluster formation, and – because it is much less affected by the complex physics of galaxies – offers a cleaner view of the underlying physical processes and fundamental cosmology. It is therefore of great interest to study the properties of the IGM, especially during the era of the first galaxies (when the IGM undergoes major changes).

One of the great triumphs of modern numerical simulations is in describing the distribution of the intergalactic matter distribution in terms of a *cosmic web* of sheets and filaments separating large voids that are nearly empty of matter (see Figure 4.1). However, the formation of these structures is actually remarkably simple, and it can be understood with a simple extension of linear perturbation theory called the *Zel'dovich approximation*[22].

Let us begin by considering the distribution of matter at a very early time t_i. We define $\mathbf{q}$ as the initial comoving position of each particle. If the universe were homogeneous, we could then write its later position as $\mathbf{r}(t) = a(t)\mathbf{q}$.

Now suppose we allow perturbations in the density field. We think of these perturbations as small displacements in the initial position of each particle, and we can express these displacements as a function of the original location, $\mathbf{p}(\mathbf{q})$. At later times, gravity will cause these displacements to change according to the local potential. As a simple approximation, let us assume that this evolution is driven entirely by the *initial* potential Φ_i. Then we can write

$$\mathbf{r}(t) = a(t)[\mathbf{q} + b(t)\mathbf{p}(\mathbf{q})], \tag{4.1}$$

where $b(t)$ is a new temporal function that describes the growth of these displacements with time. Note that because we assume that the displacement field is driven by the potential at a fixed time, the *direction* of the perturbation does not change with time, only its amplitude does. This approximation ignores the later evolution in the potential driven by these perturbations, so it represents a limited extension of perturbation theory.

The coordinates $\mathbf{q}$ are known as **Lagrangian** coordinates, because they label individual mass parcels; the Lagrangian coordinates of the parcels do not evolve with time. They are not the same as comoving coordinates $\mathbf{x}$, which are defined by $\mathbf{r}(t) = a(t)\mathbf{x}(t)$. Comoving coordinates are an **Eulerian** system, meaning they

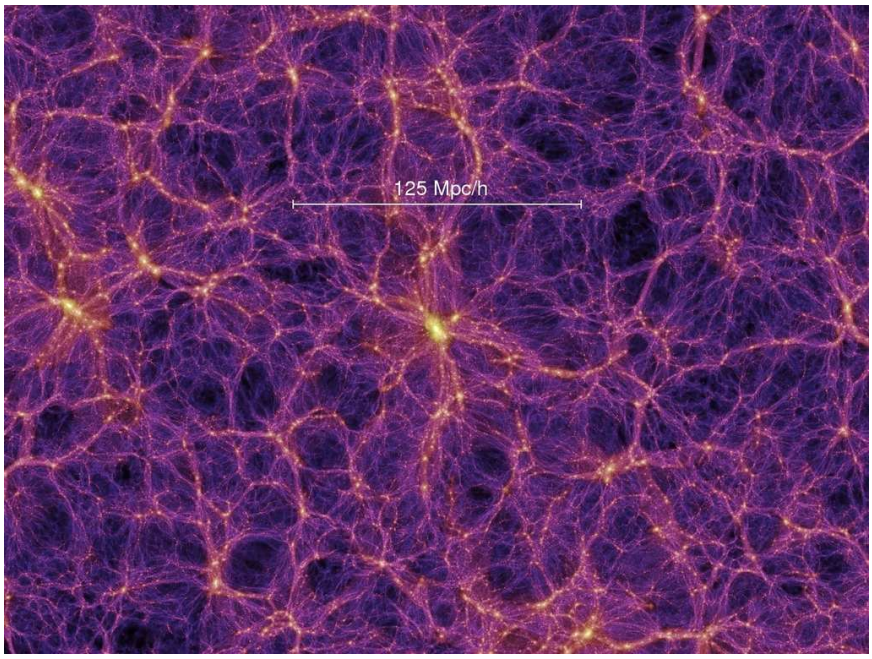

Figure 4.1 Slice through the *Millennium Simulation*, a massive computer simulation of cosmological structure formation. The colorscale shows the dark matter density; note how matter is organized into dense filaments (in many cases, these are actually slices through sheets of matter) separating nearly empty voids. Massive galaxies and galaxy clusters form at the intersections of these filaments. Figure credit: V. Springel et al. (2005).

refer to a fixed spatial grid rather than labeling particles. The comoving position can evolve as a particle moves, whereas the Lagrangian position $\mathbf{q}$ does not.

Let us now consider the evolution of the density field in the Lagrangian framework. Conservation of mass demands $\rho(\mathbf{r},t)d^3\mathbf{r} = \bar{\rho}d^3\mathbf{q}$, where in the Lagrangian system the density perturbations are contained entirely within the spacings of the coordinate grid $\mathbf{q}$. Thus, the Jacobian gives

$$\rho(\mathbf{r},t) = \bar{\rho}\det(\partial q^i/\partial r_j) \tag{4.2}$$

$$= \frac{\bar{\rho}(t)}{\det[\delta_{ij} + b(t)(\partial p_j/\partial q_i)]}, \tag{4.3}$$

or to first order in $b(t)\mathbf{p}(\mathbf{q})$, the density perturbation $\delta \equiv \rho/\bar{\rho} - 1$ is

$$\delta = -b(t)\nabla_{\mathbf{q}} \cdot \mathbf{p}, \tag{4.4}$$

where $\nabla_{\mathbf{q}}$ is the gradient with respect to the Lagrangian coordinate system.

It is convenient to Fourier transform the density field, as in equation (2.8), except we separate the time dependence of the growing mode:

$$\delta = D(t)\int \frac{d^3k}{(2\pi)^3}\delta_{\mathbf{k},i}e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{4.5}$$

where $\delta_{\mathbf{k},i}$ is the Fourier transform of $\delta(t_i)$. Fourier transforming equation (4.4) and comparing to this expression, we see first that in order for the time dependence to match we must have $b(t) = D(t)$, the normal growth factor, and then

$$\mathbf{p}(\mathbf{q}) = -i\frac{\delta_{\mathbf{k},0}}{k}\hat{\mathbf{k}}. \tag{4.6}$$

Not surprisingly, this has the same form as the peculiar velocity $\mathbf{u}$ in equation (2.10): the displacement field is simply the linear-order peculiar velocity of each particle integrated over time.

By dotting $\mathbf{k}$ into equation (4.6), it is clear that $\mathbf{p}(\mathbf{q})$ is the gradient of a function. This implies that the matrix $\partial p_j/\partial q_i$ is a real, symmetric matrix that can be diagonalized to obtain three real eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and their associated principal axes. As a result, the determinant in equation (4.3) may be written as,

$$\rho(\mathbf{r},t) = \frac{\bar{\rho}(t)}{[1-b(t)\lambda_1(\mathbf{q})][1-b(t)\lambda_2(\mathbf{q})][1-b(t)\lambda_3(\mathbf{q})]}. \tag{4.7}$$

This has a straightforward physical interpretation. Consider an infinitesimal cube surrounding each point in space and containing a set of neighboring particles. The peculiar velocities of these particles deform the cube over time. The principal axes of the transformation $\mathbf{p}(\mathbf{q})$ define the principal axes by which this cube is deformed, and the eigenvalues λ_i are proportional to the growth rate of the deformation along these axes.

When $D(t)\lambda_1 = 1$, the collection of particles has collapsed into a sheet perpendicular to the first principal axis. This approximation therefore predicts that two-dimensional "sheets" or "pancakes" will be the first nonlinear structures to form. Once collapse occurs along a second axis, a one-dimensional filament will

form, and once the third axis collapses, a halo forms.[i]

This qualitative picture matches up nicely with the cosmic web seen in numerical simulations, and indeed the Zel'dovich approximation works surprisingly well even into the nonlinear regime. There are two ways to understand this impressive success. First, the Zel'dovich approximation only requires that $b(t)\mathbf{p} \ll \mathbf{q}$. That is less restrictive than requiring $\delta \ll 1$, because δ is a function of derivatives of $\mathbf{p}$, which can get large well before the displacement field itself does. Second, it is easy to see that the Zel'dovich approximation is exact in one dimension. In that case, the gravitational dynamics just following sheets of matter, and the acceleration toward a sheet is independent of distance. Thus, in one dimension, the net acceleration experienced at a point only depends on the number of mass sheets on either side of it, which remains constant until "shell-crossing" at collapse. One can therefore extrapolate positions from the initial displacement field with the constant velocity field $b(t)\mathbf{p}$ exactly, at least until shell-crossing. To the extent that collapse along the λ_1 axis is much faster than that along the other two axes, we therefore expect the Zel'dovich approximation to describe the initial collapse very well.

4.2 LYMAN-α ABSORPTION IN THE INTERGALACTIC MEDIUM

Although dark matter dominates the mass budget of the IGM, it is the baryons which most concern us, since they provide the fuel for galaxy formation, interact with the radiation from galaxies, and – most importantly – provide observables that allow us to trace the structure of the cosmic web.

Hydrogen is the most abundant element in the Universe, making up $\approx 93\%$ of the atoms in the Universe (the remainder is almost all helium). This is now a well-understood result of the hot Big Bang model, in which nucleosynthesis (completed within the first few minutes after the Big Bang) efficiently combined all the remaining neutrons into helium atoms but then got bottlenecked by the lack of stable isotopes with 5 or 8 nucleons. As a result, all of the heavier elements (or *metals* in astronomers' parlance) were formed in the interiors of stars within galaxies. We expect (and observations confirm) that the IGM is even more dominated by hydrogen and helium than the Milky Way. We therefore focus on these two elements – and especially hydrogen – in our study of the IGM.

Since the lifetime of energy levels with principal quantum number $n > 1$ is far shorter than the typical time it takes to excite them in the rarefied environments of the Universe, hydrogen is nearly always found to be in its ground state (lowest energy level) with $n = 1$. This implies that the transitions we should focus on are those that involve the $n = 1$ state. Below we describe two such transitions, depicted in Figure 4.2.

[i] Strictly speaking, this complete collapse does not occur in the Zel'dovich approximation, because the particles continue to travel in their original direction of motion. Thus, shortly after collapse to a sheet, the particles cross each other and the sheet expands again. Obviously, the problem lies in assuming a constant peculiar velocity set by the initial potential; once collapse occurs the potential has changed significantly. The so-called *adhesion* model[23] improves the Zel'dovich approximation to account for this effect.

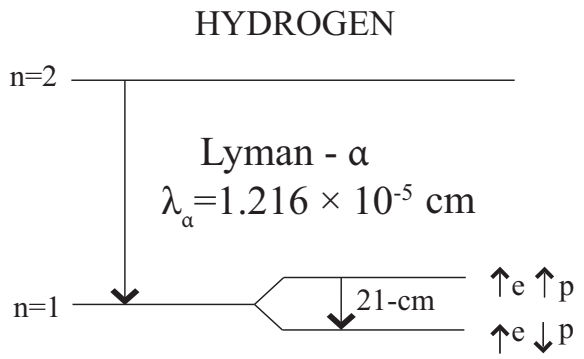

Figure 4.2 Two important transitions of the Hydrogen atom. The 21-cm transition of hydrogen is between two slightly separated (hyperfine) states of the ground energy level (principal quantum number $n = 1$). In the higher energy state, the spin of the electron (e) is aligned with that of the proton (p), and in the lower energy state the two are anti-aligned. A spin flip of the electron results in the emission of a photon with a wavelength of 21-cm (or a frequency of 1420 MHz). The second transition is between the $n = 2$ and the $n = 1$ levels, resulting in the emission of a Lyman-α photon of wavelength $\lambda_\alpha = 1.216 \times 10^{-5}$ cm (or a frequency of 2.468×10^{15} Hz).

The most widely discussed transition of hydrogen in cosmology is the Lyman-α spectral line, in which an electron moves between the $n = 1$ and $n = 2$ electronic states and which was discovered experimentally in 1905 by Harvard physicist Theodore Lyman. This line has been traditionally used to probe the ionization state of the IGM in the spectra of quasars, galaxies, and gamma-ray bursts. Back in 1965, Peter Scheuer[24] and, independently, Jim Gunn & Bruce Peterson[25] realized that the cross-section for Lyman-α absorption is so large that the IGM should be opaque to it even if its neutral (non-ionized) fraction is as small as $\sim 10^{-5}$.

Imagine a photon emitted at a wavelength $\lambda < \lambda_\alpha$, where $\lambda_\alpha = 1216$ Å is the wavelength of the Lyman-α transition. As the photon travels through the IGM, it redshifts along with the expanding Universe. Eventually, its wavelength stretches near the Lyman-α resonance, where it can be absorbed by a hydrogen atom and re-emitted in a different direction. We therefore compute the optical depth intercepted by the photon by integrating all the way across the resonance line. We will let $\lambda_{\rm obs}$ ($\nu_{\rm obs}$) be the observed wavelength (frequency).

The full cross section of a single atom is

$$\sigma_\alpha(\nu) = \frac{3\lambda_\alpha^2\Lambda_\alpha^2}{8\pi}\frac{(\nu/\nu_\alpha)^4}{4\pi^2(\nu-\nu_\alpha)^2 + (\Lambda_\alpha^2/4)(\nu/\nu_\alpha)^6}, \tag{4.8}$$

where $\Lambda_\alpha = (8\pi^2 e^2 f_\alpha/3m_e c\lambda_\alpha^2) = 6.25 \times 10^8$ s^{-1} is the Lyman-α ($2p \rightarrow 1s$) decay rate, $f_\alpha = 0.4162$ is the oscillator strength, and $\lambda_\alpha = 1216$Å and $\nu_\alpha = (c/\lambda_\alpha) = 2.47 \times 10^{15}$ Hz are the wavelength and frequency of the Lyman-α line. The term in the numerator is responsible for the classical Rayleigh scattering.

In practice, the IGM atoms have a finite spread in their thermal velocities – as well as peculiar velocities and (possibly) turbulence – all of which move the line

center around in velocity space. These mechanisms are typically modeled with a Gaussian line profile, $\sigma \propto \exp[-(\nu-\nu_\alpha)^2/2b^2]$, with the Doppler parameter b. The total absorption therefore follows the convolution of this Gaussian and the intrinsic Lorentzian profile, known as a **Voigt profile**; this is Gaussian in its core, but with much more extended Lorentzian wings.

For the moment, we will assume that the photon begins its journey with a wavelength λ much farther from resonance than the line width (see §**??** for a discussion of the more general case). We can then approximate the line as narrow,

$$\sigma_\alpha(\nu) = \frac{3\Lambda\lambda_\alpha^2}{8\pi}\delta(\nu-\nu_\alpha). \tag{4.9}$$

Then, if r is the photon's proper distance from us and the neutral hydrogen density is $n_{\rm HI}(z) = x_{\rm HI} n_H(z)$ with $x_{\rm HI}$ the neutral fraction and n_H the number density of hydrogen nuclei,

$$\tau_\alpha = \int dr\, \sigma(r) n_H(r) \tag{4.10}$$

$$= \frac{c}{H_0}\int \frac{da}{a}\sigma(\nu_{\rm obs}/a) n_H(a)[\Omega_m/a^3+\Omega_\Lambda]^{-1/2} \tag{4.11}$$

$$= \frac{3\Lambda\lambda_\alpha^3}{8\pi}\frac{x_{\rm HI} n_H(z)}{H(z)} \tag{4.12}$$

$$\approx 1.6\times 10^5 x_{\rm HI}(1+\delta)\left(\frac{1+z}{4}\right)^{3/2}. \tag{4.13}$$

where we have used $dr = cdt = cda/\dot{a} = c(da/aH)$ with the Hubble parameter $H = (\dot{a}/a)$ evaluated in the matter-dominated era. We have also let $n_H(z) = \bar{n}_H(z)(1+\delta)$ in the last line.

Obviously, the IGM optical depth can be enormous even if the neutral fraction is small. *Any* transmission across these wavelengths is therefore evidence that the diffuse IGM is highly ionized.

In practice, the IGM absorption is observed by against a luminous background source (either a bright quasar or bright gamma-ray burst afterglow). In either case, the source emits photons over an extended continuum, and we expect features in the spectrum where this Lyman-α forest begins. If the source resides at a redshift z_s, this transition will occur at an observed wavelength $\lambda_\alpha(1+z_s)$. Photons redward of this point begin their journeys at $\lambda > \lambda_\alpha$ and redshift as they travel, so they never resonate with the Lyman-α line in the IGM (though they may be absorbed by other species; see §4.5 below).

On the other hand, photons blueward of this point will eventually redshift into resonance and (if the gas is not too highly ionized) be absorbed. We therefore expect a break in the spectrum at $\lambda_\alpha(1+z_s)$, with a depth depending on the ionized fraction of the IGM and z_s (which affects the proper density of hydrogen). At moderate and high redshifts ($z > 3$), this "Lyman-break" is substantial enough to be useful as a redshift estimator. In fact, one of the premier techniques for identifying high-z galaxies is by photometrically identifying extended sources with strong flux redward of the wavelength corresponding to the seeked-after z_s and little or no flux blueward of that wavelength. Note that by $z \sim 3$ this break has

redshifted into the optical range and is easy to observe from the ground; by $z \sim 6$ it has redshifted into the near-infrared, where observations are more challenging.

Naively, how would one expect the optical depth to evolve? The redshift factor $(1+z)^{3/2}$ reflects the evolution of the column density of hydrogen atoms, and implies a slow increase for τ_α as redshift increases. But more important is the factor $x_{\rm HI}$, which will evolve both with the background density and the ionizing background. As redshift increases, one might naturally expect the number of ionizing sources to decrease, because structure formation is less advanced. In that case we would expect the optical depth to increase even faster, with the IGM eventually becoming opaque once the ionizing background falls far enough. (In practice, the ionizing background appears to be roughly constant with redshift at $3 < z < 6$, but it must eventually decrease at higher redshifts during or before "reionization.")

If a source were to be observed when the atomic fraction of hydrogen were substantial, then *all* photons with wavelengths just short of the Lyman-α wavelength at the source (observed at $1216(1+z_s)$Å, where z_s is the source redshift) would redshift into resonance, be absorbed by the IGM, and then get re-emitted in other directions. Eventually, this would result in an observed *complete* absorption trough shortward of λ_α in the source spectrum, known as the "Gunn-Peterson trough."

Figure 4.3 shows spectra of 19 quasars at $z \sim 6$; note how indeed the fraction of transmitted flux blueward of the Lyman-α line of each quasar decreases toward the higher redshifts in this range. The spectra of the highest-redshift quasars at $z < 6.4$ show hints of a Gunn-Peterson effect. Unfortunately, this is difficult to interpret because only a very small neutral fraction is required to saturate the Gunn-Peterson trough (see equation 4.13). We cannot yet determine whether the IGM is slightly ionized or nearly neutral at this time.

Figure 4.3 also shows that as the forest becomes clearer the absorption is also highly non-uniform. This becomes even more pronounced at moderate redshifts, as shown by the example in Figure **??**. We now understand this "forest" of features to originate from the cosmic web: as a line-of-sight passes through the sheets, filaments, and voids of the cosmic web, the optical depth fluctuates, creating the so-called "Lyman-α forest" of absorption features. Because each observed wavelength corresponds to λ_α at a specific distance from us, each of the absorption features can be associated with a single feature in the density field: photons that hit the Lyman-α resonance at precisely that point may get absorbed, but others pass through without interacting. It is this forest that provides most of our knowledge about the IGM at moderate and low redshifts, and we will next study the physics behind it.

4.3 THEORETICAL MODELS OF THE LYMAN-α FOREST

To compute the optical depth distribution of the IGM, it is therefore necessary to know how the neutral fraction $x_{\rm HI}$ varies through space. To a very good approximation, almost all regions are in *ionization equilibrium*, where the number of ionizations per second balances the number of recombinations,

$$n_e n_p \alpha(T) = n_{\rm HI}\Gamma \tag{4.14}$$

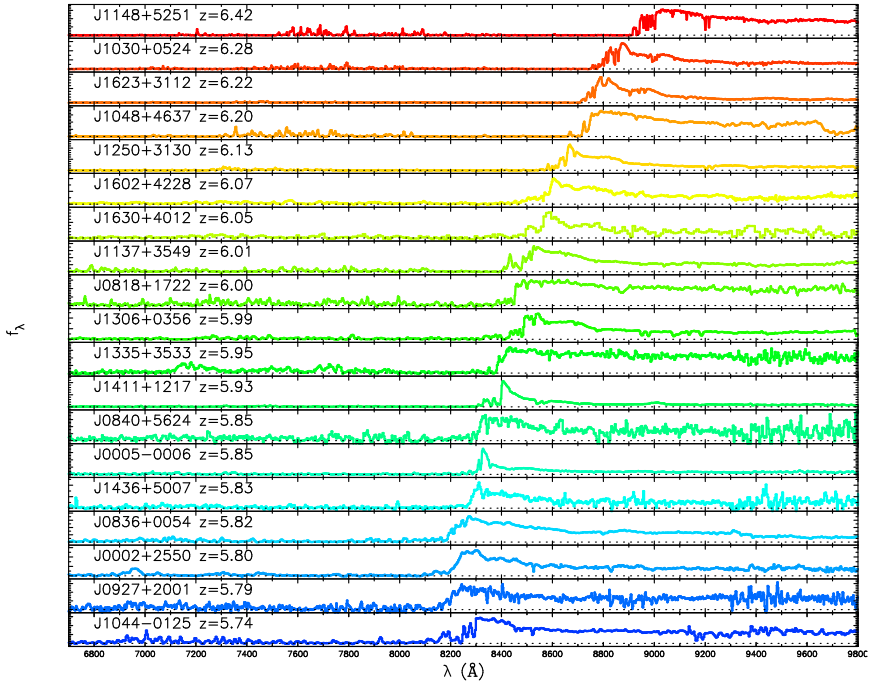

Figure 4.3 Observed spectra (flux per unit wavelength) of 19 quasars with redshifts $5.74 < z < 6.42$ from the Sloan Digital Sky Survey. For some of the highest-redshift quasars, the spectrum shows no transmitted flux shortward of the Lyman-α wavelength at the quasar redshift, providing a possible hint of the so-called " Gunn-Peterson trough" and indicating a slightly increased neutral fraction of the IGM. It is evident from these spectra that broad-band photometry is adequate for inferring the redshift of sources during the epoch of reionization. Figure credit: Fan, X., et al. *Astron. J.* **128**, 515 (2004).

where $\alpha(T)$ is the (temperature-dependent) recombination coefficient and the ionization rate (per atom) is

$$\Gamma = \int_{\nu_L}^{\infty} d\nu \, \frac{4\pi J(\nu)\sigma_{\rm HI}(\nu)}{h\nu} \tag{4.15}$$

with $J(\nu)$ being the specific intensity of the background field (in units erg cm^{-2} s^{-1} Hz^{-1} sr^{-1}) and $\sigma_{\rm HI}$ is the cross section for ionization. This integral counts the number of photons per second striking an atom, weighted by the ionization cross section. As we will see below, typical values for the ionization rate are $\Gamma \approx 10^{-12}$ s^{-1}, and so we will normalize $\Gamma = \Gamma_{12} \times 10^{-12}$ s^{-1} for convenience.

The photo-ionization cross-section applies to wavelengths shorter than $\lambda_c = 912$Å, at which hydrogen or helium are ionized. The bound-free absorption cross-section from the ground state of a hydrogenic ion with nuclear charge Z and an ionization threshold $h\nu_0$ is given by,

$$\sigma_{bf}(\nu) = \frac{6.30 \times 10^{-18}}{Z^2} \, {\rm cm}^2 \times \left(\frac{\nu_0}{\nu}\right)^4 \frac{e^{4-(4\tan^{-1}\epsilon)/\epsilon}}{1-e^{-2\pi/\epsilon}} \quad \text{for } \nu \geq \nu_0, \tag{4.16}$$

where

$$\epsilon \equiv \sqrt{\frac{\nu}{\nu_0} - 1}. \tag{4.17}$$

For neutral hydrogen, $Z = 1$ and $\nu_{\rm H,0} = (c/\lambda_c) = 3.29 \times 10^{15}$ Hz ($h\nu_{\rm H,0} = 13.60$ eV); for singly-ionized helium, $Z = 2$ and $\nu_{\rm He\ II,0} = 1.31 \times 10^{16}$ Hz ($h\nu_{\rm He\ II,0} = 54.42$ eV).

A simpler fitting formula for these cross-sections commonly used in numerical calculations is

$$\sigma_{\rm HI,HeII}(\nu) = \sigma_0 (E_* - 1)^2 \frac{E_*^{-4.02}}{(1+\sqrt{E_*/32.88})^{2.963}}, \tag{4.18}$$

where $E_\star = E/E_0$. Here $E_0 = 0.4298$ eV and $\sigma_0 = 5.475 \times 10^{-14}$ cm^2 for hydrogen and $E_0 = 1.720$ eV and $\sigma_0 = 1.369 \times 10^{-14}$ cm^2 for singly-ionized helium. Although hardly obvious, this follows $\sigma_{\rm HI} \propto \nu^{-3}$ near the ionization threshold.

The cross-section for neutral helium is more complicated; when averaged over its narrow resonances it can be fitted to an accuracy of a few percent up to $h\nu = 50$ keV by the fitting function [26],

$$\sigma_{bf,{\rm He\ I}}(\nu) = 9.492 \times 10^{-16} \, {\rm cm}^2 \times \left[(x-1)^2 + 4.158\right] \times y^{-1.953} \left(1 + 0.825 y^{1/4}\right)^{-3.188}, \tag{4.19}$$

where $x \equiv [(\nu/3.286 \times 10^{15} \, {\rm Hz}) - 0.4434]$, $y \equiv x^2 + 4.563$, and the threshold for ionization is $\nu_{\rm He\ I,0} = 5.938 \times 10^{15}$ Hz ($h\nu_{\rm He\ I,0} = 24.59$ eV).

The recombination coefficient $\alpha(T)$ describes the rate at which electrons and protons recombine (while emitting a photon). Of course, the recombination can occur to any of the hydrogen atom's energy levels; of particular interest is recombination to the ground state, which generates a new ionizing photon. Provided that

photon is absorbed by the gas, such a recombination does not lead to a net increase in the neutral fraction. The same is true for resonant photons, such as Lyman-α, which through repeated emission and absorption events do not lead to net recombinations in a gas that is optically-thick to them. It is therefore often useful to consider the "case-B" recombination coefficient α_B, which excludes recombinations to the ground state. For hydrogen,

$$\alpha_B(T) \approx 2.6 \times 10^{-13} T_4^{-0.76} \text{ cm}^3 \text{ s}^{-1}, \tag{4.20}$$

where $T_4 = T/10^4$ K. The contrasting case, where such photons escape the region of interest, is referred to as "case-A" and has a rate coefficient

$$\alpha_A(T) \approx 4.2 \times 10^{-13} T_4^{-0.76} \text{ cm}^3 \text{ s}^{-1}. \tag{4.21}$$

Note that both rates are fairly slow in the IGM, except at high redshifts. At the mean density of the IGM, the ratio of the case-B recombination time, $t^B_{\rm rec} = 1/n_H\alpha$, to the age of the Universe is

$$\frac{t^B_{\rm rec}}{t_H} \approx 0.8 \left(\frac{8}{1+z}\right)^{3/2}. \tag{4.22}$$

In other words, once an atom is ionized at $z < 7$, it may remain ionized forever at the mean cosmic density.

The appropriate coefficient to use depends on the physical situation at hand. If one is concerned with the average absorption in a uniform IGM, case-B is clearly the best choice, because photons from recombinations to the ground state will be absorbed somewhere in the IGM. If, on the other hand, the IGM is very clumpy so that most of the recombination photons will be absorbed inside dense neutral blobs, case-A is a better choice. Similarly, if one considers ionization equilibrium in a single dense cloud, case-A may be more appropriate if the recombinations occur preferentially on the "skin" of the cloud, so that the resulting photons can easily escape to the external medium.

In the highly ionized limit of equation (4.14), we can equate n_p to the total proton density; in that case, using the case-B recombination rate,

$$x_{\rm HI} = n_e \alpha(T) \Gamma^{-1} \sim 4 \times 10^{-6} (1+\delta) \left(\frac{1+z}{4}\right)^3 T_4^{-0.76} \Gamma_{12}^{-1}. \tag{4.23}$$

Note that, for detailed calculations we should include the electrons from ionized helium, but that makes only a minor difference at the level of $\sim 10\%$.

Clearly the gas is indeed highly ionized, at least at moderate and low redshifts. Conveniently, substituting this value into equation (4.13), the optical depth for gas at the mean density (and $z \sim 3$) is of order unity, just in the range in which we can accurately measure the absorption. The Lyman-α forest therefore allows us to map the cosmic web in exquisite detail, even though the gas itself is at extremely low densities.

4.3.1 The Temperature-Density Relation

To proceed farther and evaluate τ_α as a function of density, we must determine the temperature of the IGM gas. Thermal equilibrium is typically established by three

competing effects. First, the adiabatic expansion of the Universe cools the gas, as does Compton cooling (which is important at $z > 6$; at lower redshifts it becomes unimportant relative to the expansion cooling). Other mechanisms – such as line cooling – are much less efficient.

On the other hand, heat input is provided by photoionization. A typical ionizing photon has more energy than the ionization threshold, so the free electron is left with residual kinetic energy. The electron then scatters through the IGM and deposits its energy as heat. The heating rate for HI (in erg cm^{-3} s^{-1}) is

$$\left.\frac{dQ}{dt}\right|_i = n_{\rm HI} \int_{\nu_i}^{\infty} d\nu \, (4\pi J_\nu) \sigma_{\rm HI}(\nu) \left(\frac{h\nu - h\nu_i}{h\nu}\right). \tag{4.24}$$

(Note that it is often important to include helium here, as it remains in its singly-ionized state until $z \sim 3$ and efficiently absorbs high-energy photons, but for pedagogical simplicity we will ignore it.)

Clearly, the photoheating rate depends sensitively on the ionized fraction of the gas. If we imagine that a gas parcel is initially neutral and then is rapidly exposed to a strong ionizing background, all of the gas will quickly be ionized. The *total* energy input is therefore simply the average excess energy per ionization $\langle E_i \rangle$, yielding

$$k_B \Delta T = \frac{2}{3} \frac{n_H}{n_{\rm tot}} \langle E_i \rangle . \tag{4.25}$$

This deceptively simple expression actually hides a fair amount of physics in the factor $\langle E_i \rangle$, which depends on how the spectrum of incident radiation interacts with the gas parcel. Two limits are illuminating. First, if the parcel is optically thin, then the weighting by $\sigma_{\rm HI}$ in equation (4.24) reduces the impact of high-energy photons. In this case,

$$\langle E_{i,\rm thin} \rangle = \frac{1}{\Gamma} \int_{\nu_i}^{\infty} d\nu \, (4\pi J_\nu) \sigma_{\rm HI}(\nu) \left(\frac{h\nu - h\nu_i}{h\nu}\right). \tag{4.26}$$

On the other hand, if the element is optically thick up to some maximum frequency (because $\sigma_{\rm HI} \propto \nu^{-3}$), then all photons below this frequency are absorbed and the weighting by $\sigma_{\rm HI}$ disappears.

The latter case can make a significant difference to the total temperature increase: for a specific luminosity $L_\nu \propto \nu^{-\alpha}$, it yields $\langle E_{i,\rm thin} \rangle = E_{\rm HI}/(\alpha + 2)$, where $E_{\rm HI} = 13.6\,$eV is the HI ionization potential. In the particular case of $L_\nu \propto \nu^{-2}$ appropriate for a low-metallicity galaxy, $\langle E_{i,\rm thin} \rangle / E_{\rm HI} \approx 1/4$. In the optically thick limit, we have instead $\langle E_{i,\rm thick} \rangle = E_{\rm HI}/(\alpha - 1)$. For $L_\nu \propto \nu^{-2}$, this yields $\langle E_{i,\rm thick} \rangle / E_{\rm HI} \approx 3/5$. The net temperature change is then $\Delta T \approx 0.5(2/3k_B) \langle E \rangle \sim 30,000\,$K for the optically-thick case, significantly above the value of $\sim 12,500\,$K for the optically-thin case.

Because this energy input is identical for each particle (modulo the optical depth of its environment), the temperature of a parcel should be *independent* of its density immediately after ionization.[ii] However, after this initial phase of ionization,

[ii] We will see later, however, that there is *on average* a non-trivial temperature-density relation during reionization, because the *cosmic time* at which elements are ionized depends on the density.

dQ/dt decreases dramatically, because $n_{\rm HI}$ becomes very small. Thus, the cooling term begins to dominate. Expanding regions that are ionized early, cool initially to lower temperatures than similar regions that are ionized later.

The resulting equilibrium *does* depend on density, because the adiabatic cooling rate depends on the expansion rate, which in turn depends on the local density – underdense voids can be considered (locally) to have a smaller Ω_m, and so they expand faster. Thus, the low-density regions cool fastest. Eventually, an equilibrium is reached in which

$$T \approx T_0(1+\delta)^{\gamma-1}, \tag{4.27}$$

with T_0 being a normalization constant and $\gamma \approx 1.6$ long after reionization. The normalization of this *temperature-density relation*[iii] is entirely determined by the *spectral shape* of the ionizing background and is independent of its amplitude, because the heating rate per neutral atom is proportional to J_ν (equation 4.24), but the neutral fraction is proportional to $1/J_\nu$ (through Γ in equation 4.23).

4.3.2 The Fluctuating Gunn-Peterson Approximation

A simple model for the absorption pattern of the inhomogeneous IGM simply associates each gas element with its "local" Gunn-Peterson optical depth in equation (4.13). This is an oversimplification for two reasons: first, it ignores the frequency structure of the line (so that the total τ_α is distributed amongst many neighboring gas elements), and second, it ignores the velocity structure of the IGM which moves gas elements around in frequency space. Nevertheless, it provides a simple description and a reasonable approximation to the parameter dependencies of the real Lyman-α forest.

With the assumption of ionization equilibrium (equation 4.23) and using the approximate power-law form of the temperature-density relation (equation 4.27), equation (4.13) becomes

$$\tau_\alpha(\delta,T) \approx 13\frac{(1+\delta)^{2-0.76(\gamma-1)}}{\Gamma_{12}}\left(\frac{T_0}{10^4\,{\rm K}}\right)^{-0.76}\left(\frac{1+z}{7}\right)^{9/2}. \tag{4.28}$$

The $(1+\delta)$ exponent ranges from 2 (for isothermal gas) to ~ 1.5 (at the thermal asymptote); it is greater than unity because of the recombination rate scaling (which also induces the temperature dependence).

Equation (4.28) shows that at $z \sim 6$ only the most underdense regions will be visible (with $\tau_\alpha < 1$); gas at the mean density will be extremely opaque even if the ionizing background is substantial. This explains the deep absorption troughs in Figure 4.3. However, at $z \sim 3$ the same gas parcel at the mean density has $\tau_\alpha \sim 1$: this is why the Lyman-α forest is such a powerful tool at moderate and low redshifts.

Because τ_α depends only on fundamental cosmological parameters (known reasonably well), the density and temperature of the IGM (which can be modeled relatively well), and the unknown Γ, the transmission in the Lyman-α forest provides

[iii] This relation is sometimes referred to as the "IGM equation of state," but that is a misnomer because the relation describes the relation averaged over different gas parcels, rather than the evolution of a single one.

a good estimate of Γ. Measurements show that $\Gamma_{12} \sim 1$ over the range $z \sim 2$–5, with uncertainties between different methods and samples of no more than a factor of ~ 2.

There is one other important property of the intergalactic gas: for a fixed average H I density, the transmission through an inhomogeneous medium is *smaller* than through a homogeneous one. To see this, let us define $p(\delta)$ as the volume-averaged probability distribution of the IGM density. Then the net transmission is

$$T_\alpha = \int d\delta\, p(\delta) \exp[-\tau_\alpha(\delta)] \tag{4.29}$$

$$\equiv \exp(-\tau_{\rm eff,\alpha}), \tag{4.30}$$

where we have defined the **effective optical depth** in the line as $\tau_{\rm eff,\alpha}$. This effective value must be smaller than the corresponding Gunn-Peterson absorption $\tau_\alpha(\delta = 0)$ because of the well-known triangle inequality,

$$\langle \exp(-\tau_\alpha) \rangle \geq \exp(-\langle \tau_\alpha \rangle). \tag{4.31}$$

Essentially, because the absorption saturates in dense regions, an inhomogeneous medium has *less* overall absorption than a uniform medium. Most of the transmission arises in the low-density voids, which can remain transparent even if the gas at the mean density is optically thick.

4.3.3 The Column Density Distribution

The fluctuating Gunn-Peterson approximation is a useful model partly because it suggests that the IGM optical depth varies continuously along the line-of-sight, just as the density field of the cosmic web does. However, in practice the Lyman-α forest appears as a set of discrete absorbers, because IGM density peaks (intercepted sheets and filaments) are rather sharp. Thus, it is often useful to consider such systems as discrete absorbers.

We begin by assuming that the absorption by a given region will be dominated by its densest portion (with peak fractional overdensity δ). In order to compute the optical depth, we must assign this region a length scale. The most natural size is the **local Jeans length**, which is simply the length scale over which the pressure force balances gravity (see §3.2), $L_{\rm J} \sim c_s t_{\rm coll} \sim c_s (G\rho)^{-1/2}$: a smaller cloud (at the same density ρ) will be smoothed out by pressure whereas a larger cloud will collapse gravitationally. If we assume that the gas maintains photoionization equilibrium, the corresponding column density through the cloud is $N_{\rm HI} = x_{\rm HI} n_H L_{\rm J}$, or

$$N_{\rm HI} = 3.3 \times 10^{14}\,{\rm cm}^{-2} (1+\delta)^{3/2} \left(\frac{T_0}{10^4\,{\rm K}}\right)^{-0.26} \Gamma_{12}^{-1} \left(\frac{1+z}{7}\right)^{9/2}. \tag{4.32}$$

As described above, the properties of these regions are measured from their optical depth for Lyman-α absorption, $\tau_\alpha(\nu) = N_{\rm HI}\sigma_\alpha(\nu)$. Thus, *in principle* the column density is an observable in the forest. In practice, $N_{\rm HI}$ can be difficult to extract from the total amount of absorption because of saturation. The **equivalent width** parameterizes the amount of absorption as the wavelength interval over

which light would be absent if the line profile were a step function,

$$W = \int [1 - e^{-\tau(\lambda)}] d\lambda. \tag{4.33}$$

When τ is small, $W \propto \int d\lambda \tau(\lambda) \propto N_{\rm HI}$. When $\tau \gg 1$, the line center is strongly saturated but the Lorentzian wings have substantial optical depth and dominate the total absorption; in that case, $W \propto N_{\rm HI}^{1/2}$. Unfortunately, in the intermediate regime where the line center is saturated but the wings remain weak, $W \propto \ln N_{\rm HI}$, and it is very difficult to measure the true column density of a line.[iv] This intermediate range approximately spans HI column densities of 10^{14}–10^{17} cm^{-2}.

Nevertheless, the physical importance of $N_{\rm HI}$ makes it the preferred parameter for describing the Lyman-α forest. Because there is a one-to-one correspondence between $N_{\rm HI}$ and the local IGM density (at least in this local Jeans length approximation), it is convenient to describe the Lyman-α forest via the number density of absorbers in a column density interval $(N_{\rm HI}, N_{\rm HI} + dN_{\rm HI})$ and in a redshift interval $(z, z + dz)$, $d^2N/dN_{\rm HI}dz$.[v]

If this distribution function is known, one can estimate the *total* optical depth $\tau_{{\rm eff},\alpha}$ in the forest by integrating over all the lines,

$$\tau_{{\rm eff},\alpha} = \frac{(1+z)}{\lambda_\alpha} \int dN_{\rm HI} \frac{d^2N}{dN_{\rm HI}dz} W(N_{\rm HI}). \tag{4.34}$$

Note that $\tau_{{\rm eff},\alpha}$ is *not* simply the average of the optical depths of all the lines, because the observed transmission depends exponentially on τ_α. Importantly, however, $\tau_{{\rm eff},\alpha}$ does not *require* a measurement of $dN/dN_{\rm HI}dz$: as the total absorption, it can be estimated even from low-resolution measurements or when the forest is so thick that Lyman-α absorbers cannot be separated.

Observations show that at $z < 5.5$,

$$\tau_{{\rm eff},\alpha} = (0.85 \pm 0.06) \left(\frac{1+z}{5}\right)^{4.3\pm0.3}. \tag{4.35}$$

At $z > 5.5$ the optical depth appears to increase even more rapidly and is nearly saturated at $z \sim 6$; we discuss this regime in §4.6.

4.3.4 Mapping the Cosmic Web

As described previously, the forest is the premier tool for measuring the properties of the IGM at $z < 5$, because it provides such a detailed view of the structures. The only drawback is the relative dearth of background sources against which absorption can be measured: "bright" quasars or GRB afterglows are rare, so to date almost all of the information has come from studying a small number of individual one-dimensional skewers of the cosmic web.

This leads to an important caveat for Lyman-α forest studies of the high-z Universe: although detailed structures are visible along the line-of-sight, inverting

[iv] The problem is ameliorated somewhat for higher Lyman series lines, which have smaller optical depths, but these lines face other challenges.

[v] Note that in practice the column density distribution is often reported in the literature relative to the coordinate X, where $dX/dz = H_0(1+z)^2/H(z)$.

these to obtain the three-dimensional structure is difficult because of **aliasing**. This refers to the possibility of random arrangements of small-scale oscillations inclined to the line-of-sight mimicking large-scale oscillations along the line-of-sight; for example, if the crests of two k-modes are aligned with the plane of the sky (but at a wide radial separation) and intersect the Lyman-α forest skewer, they would appear to an observer as two crests of a single, large-wavelength oscillation along the line-of-sight.

To quantify the importance of aliasing, we begin with the correlation function: statistical isotropy guarantees that it is identical in every direction and so can be measured with data along only the line-of-sight. It is related to the three-dimensional power spectrum $P_{\rm 3D}$ through a Fourier transformation (equation 2.15). However, if we use the Lyman-α forest data itself to measure a power spectrum, we obtain only a one-dimensional power spectrum, $P_{\rm 1D}$. This is *not* the same as P_{3D}, as the following argument illustrates. Let $k_{\parallel}$ and x be the wavenumber and distance coordinate along the line-of-sight. Then

$$P_{\rm 1D}(k_1) = \int dx\, \xi(x) e^{ik_1 x} \tag{4.36}$$

$$= \int dx e^{ik_1 x} \int \frac{d^3k}{(2\pi)^3} P_{\rm 3D}(k) e^{-ikx}. \tag{4.37}$$

Note that, because x is along the line-of-sight, the y and z coordinates vanish in the second exponential. Now, integrating over x yields a factor $2\pi\delta(k-k_1)$, and implies that

$$P_{\rm 1D}(k_1) = \int \frac{dk_y dk_z}{(2\pi)^2} P_{\rm 3D}(\sqrt{k_1^2 + k_y^2 + k_z^2}) \tag{4.38}$$

$$= \int_{|k_1|}^{\infty} \frac{dk}{2\pi}\, k P_{\rm 3D}(k), \tag{4.39}$$

where we have simplified the integral by transforming to polar coordinates. This form shows the difficulty in measuring long-wavelength modes: the observed one-dimensional power at a scale k_1 picks up contributions from *all* wavenumbers greater than this value – and weighted toward the high-k contribution: if $P_{\rm 3D} \propto k^{-n}$, then the observed $P_{\rm 1D} \propto k^{2-n}$.

Thus, the Lyman-α forest is best at constraining cosmological information on small physical scales. Of course, it is precisely these scales that are most difficult to model, so numerical simulations are necessary for quantitative constraints. This procedure also helps to constrain astrophysical parameters that affect the forest – most importantly, the ionizing background (which sets the overall normalization) and the temperature (which sets the maximum wavenumber of interest through thermal broadening and Jeans smoothing of the IGM features). To date, the constraints on the ionizing background are comparable with other methods, as are the temperature measurements (although the latter have very large errors in practice).

With the advent of large-scale, deep surveys, there are now plans to observe a dense array of skewers associated with a large number of quasars and map the related large-scale structure in three dimensions. This exciting prospect will provide much better constraints on the sought-after baryon acoustic oscillations which appear on very large scales.

4.4 THE METAGALACTIC IONIZING BACKGROUND

Presuming that one can model the structure of the IGM reliably, the primary physical input determining the opacity of the Lyman-α forest is the ionization rate Γ, which in turn depends on the angle-averaged specific intensity of the radiation background, $J(\nu)$ (equation 4.15). We assume a constant emissivity $\epsilon(\nu)$ (in units of $\mathrm{erg\ s^{-1}\ cm^{-3}}$), defined at a frequency ν and redshift z. The optical depth experienced by an ionizing photon is proportional to its path length r through the IGM, $\tau(\nu, z) = r/\lambda(\nu, z)$, where $\lambda(\nu, z)$ is the mean-free-path or attenuation length of photons with a frequency ν at a redshift z. Then we have

$$J(\nu, z) = \int_0^\infty 4\pi r^2 dr \frac{\epsilon(\nu, z)}{(4\pi r)^2} e^{-r/\lambda(\nu,z)} = \frac{1}{4\pi}\epsilon(\nu, z)\lambda(\nu, z), \tag{4.40}$$

Here we have assumed that λ is much smaller than the Hubble length, so that evolution in the source density and redshifting of the photons is negligible. This is a reasonable approximation at high redshifts, except for the highest energy photons. When that is not true, one must be sure to evaluate the emissivity and optical depth at the appropriate redshift.

The emissivity clearly depends only on the sources – galaxies and quasars – and understanding this coefficient will be a key goal of the following chapters. In brief, stellar sources typically have relatively soft spectra: hot stars with a surface temperature $\sim 30,000\,\mathrm{K}$, for example, have their blackbody peak at $E \sim 7\,\mathrm{eV}$ with their emission luminosity declining sharply at higher photon energies. The spectrum of solar-mass stars cuts off well before the Lyman-limit and does not contribute significantly to the ionizing photon budget. Thus, because hot massive stars have such short lifetimes, only actively star-forming galaxies contribute to the metagalactic background. However, even their photons must escape absorption by the gas and dust inside their interstellar medium; this appears to be a difficult step in most known galaxy populations where the so-called **escape fraction** is only a few percent. It is therefore difficult to estimate the net emissivity from galaxies.

Quasars, the second important class of sources, are somewhat easier to model partly because they are brighter and hence easier to characterize. Their intense, high-energy radiation fields – with typically power-law spectra extending to X-ray energies – produce many more ionizing photons per unit energy output and probably allow a much larger fraction of these photons to escape to the IGM. In practice, both kinds of sources appear to be important at moderate redshifts. However, beyond $z \sim 4$ the bright quasar population begins to decline precipitously while the comoving star formation rate remains similar in magnitude. The natural expectation is therefore that galaxies become increasingly important at high redshifts.

The mean-free-path λ is determined by absorption in the IGM – and hence the Lyman-α forest. We next consider how to estimate this factor.

4.4.1 The Mean Free Path of Ionizing Photons

The total opacity per unit redshift of the IGM at a frequency ν is just the sum of the opacity of all the individual absorbers,

$$\frac{d\tau_{\rm HI}(\nu)}{dz} = \int dN_{\rm HI}\, \frac{d^2N}{dN_{\rm HI}dz}[1-\exp\tau(\nu,N_{\rm HI})], \tag{4.41}$$

where the optical depth of an absorber to ionizing photons is $\tau(\nu,N_{\rm HI}) = N_{\rm HI}\sigma_{\rm HI}(\nu)$. To estimate the mean-free-path, we simply convert this to a comoving path length:

$$\lambda(\nu,z) = \frac{dr/dz}{d\tau_{\rm HI}(\nu)/dz}, \tag{4.42}$$

where dr/dz is the comoving line element.

Given the distribution function of Lyman-α absorbers, this is a well-posed calculation, so the mean-free-path might appear to be straightforward to predict from first principles. However, recall that τ_α is itself a function of Γ, which in turn depends upon the mean-free-path. Self-consistently predicting the attenuation – and with it the ionizing background – is therefore a rather complex problem.

To understand better the nature of the absorbing systems, it is useful to consider the opacity as a function of column density. At the ionization threshold, $\tau_{\rm HI} = 1$ for $N_{\rm HI} = 1/\sigma_{\rm HI}(\nu_{\rm HI}) = 1.6\times10^{17}$ cm^{-2}. Systems above this column density limit are opaque to ionizing photons; we refer to this regime as **self-shielding** and these opaque systems as **Lyman-limit systems** (LLSs). The former suggests that gas on the outskirts of the system absorbs a large fraction of the incident ionizing background, shielding the interior from ionizing photons.

Does most of the opacity originate from these opaque systems or from the accumulated opacity of lower column density systems? Let us suppose for simplicity that $d^2N/dN_{\rm HI}dz = A[N_{\rm HI}\sigma_{\rm HI}(\nu_{\rm HI})]^{-\beta} \approx A\tau^{-\beta}(\nu/\nu_{\rm HI})^{-3\beta}$, where τ is defined at the ionization threshold; a single power law with $\beta\approx3/2$ provides a reasonable, though not perfect, approximation to the observed distribution (see below). The mean-free-path is then (for $1\le\beta\le2$)

$$\lambda(\nu) = \left[\frac{A}{\sigma_{\rm HI}(\nu_{\rm HI})}\right]^{-1}\left(\frac{\nu}{\nu_{\rm HI}}\right)^{-3(1-\beta)}\left[\int_0^\infty d\tau\,\tau^{-\beta}(1-e^{-\tau})\right]^{-1} \tag{4.43}$$

$$\approx \frac{1}{\Gamma_G(2-\beta)}\lambda_{\rm LLS}\left(\frac{\nu}{\nu_{\rm HI}}\right)^{-3(1-\beta)}. \tag{4.44}$$

Here Γ_G is the Gamma function (not to be confused with the ionization rate) and we have assumed that the absorbers span the range from $\tau\ll1$ to $\tau\gg1$ (with a single power law), and $\lambda_{\rm LLS}$ is the mean-free-path at the ionization edge including absorption only from systems with $\tau>1$ (we normalize to this value because it is relatively easy to measure). For $\beta=1.5$, which provides a reasonable match to the Lyman-α forest, $\lambda(\nu_{\rm HI})\approx0.56\lambda_{\rm LLS}$. Thus $\sim56\%$ of the absorption comes from the opaque systems, although this fraction is reasonably sensitive to the precise shape of the column density distribution. The evolution and distribution of these LLSs is therefore crucial to understanding the ionizing background.

Clearly, the mean-free-path is much longer for high-energy photons: with the canonical value $\beta=3/2$ we have $\lambda\propto\nu^{3/2}$. Note, however, that this is much

weaker dependence than the $\lambda \propto \nu^3$ dependence expected in a uniform IGM; the clumps are very important for high-energy photons.

Although the self-shielded absorbers are opaque to photons at the ionization edge, the strong frequency dependence of the ionization cross-section implies that they are transparent to higher-energy photons. Typically, this implies that LLSs are themselves highly-ionized. If we assume that a system with column density $N_{\rm HI}$ is opaque to all photons with $\nu < \nu_{\rm min}$ and that $\epsilon(\nu) \propto \nu^{-\alpha}$, we have $\Gamma \propto [N_{\rm HI}\sigma_{\rm HI}(\nu_{\rm HI})]^{(-\alpha+3\beta-6)/3}$, so according to equation (4.23) the residual neutral fraction is

$$x_{\rm HI} \sim 2.2 \times 10^{-4} T_4^{-0.59} \Gamma_{12}^{-1/3} [N_{\rm HI}\sigma_{\rm HI}(\nu_{\rm HI})]^{(-\alpha+3\beta-6)/3}. \quad (4.45)$$

Here, we have set our fiducial value of δ to match that of a LLS (at the ionization edge) at $z \sim 3$ using the relation

$$1 + \delta_{\rm LLS} = 320\, T_4^{0.17} \Gamma_{12}^{2/3} \left(\frac{1+z}{4}\right)^{-3}. \quad (4.46)$$

4.4.2 Observations of Lyman-Limit Systems

We have already described the important role of LLSs in setting the mean-free path of ionizing photons and hence regulating the ionizing background. Conveniently, these systems are relatively easy to identify even at high redshifts, because their optical thickness to ionizing photons causes a continuum depression in the background source's flux blueward of 912 Å in the rest frame of the absorber. Thus, these systems constitute the one family of hydrogen absorbers whose abundance at $z > 5$ has been measured.

Recent surveys have established the LLS abundance reasonably well at $0 < z < 6$; the additional assumption that $d^2N/dN_{\rm HI}dz \propto N_{\rm HI}^{-\beta}$ with $\beta \approx 1.1$–1.5 (and constant with redshift) yields a mean-free-path at the Lyman edge of[vi]

$$\lambda(\nu_{\rm HI}) \approx (50 \pm 10) \left(\frac{1+z}{4.5}\right)^{-4.44\pm0.3} \text{ proper Mpc}. \quad (4.47)$$

Interestingly, extrapolation to $z \sim 6$ (10) yields a mean-free-path of ~ 7 (1) proper Mpc – clearly, during the era of the first galaxies, ionizing photons suffer *much* more attenuation than at later times. This obviously leads to strong fluctuations in the ionizing background itself and substantially affects the process of reionization. We will consider the importance of LLS in more detail later: in essence, they regulate the end of reionization and provide the "matching condition" from the epoch of reionization to later times.

Despite the relative ease of *finding* LLSs, their physical nature remains obscure. Equation (4.45) shows that, at moderate redshifts, these objects have overdensities comparable to those inside virialized halos. As such, they are difficult to model, requiring high-resolution numerical simulations of the structure of gas around galaxies, coupled with a large enough cosmic volume to represent adequately the cosmic

[vi] Songaila & Cowie

radiation field. Explanations for their origin range from low-mass dark matter halos without substantial star formation to cold gas accreting onto galactic halos from filaments in the cosmic web.

To complicate matters further, the very nature of these LLSs may evolve at higher redshifts. Even assuming optimistically that the ionizing background remains constant, equation (4.45) shows that $\delta_{\rm LLS} \sim 20$ at $z \sim 10$.

Absorbers with extremely high column densities ($N_{\rm HI} > 10^{20.3}$ cm^{-2}) have prominent damping wings from natural line broadening and are known as **damped Lyman-α absorbers** (DLAs). Such large columns are opaque to photons with $E < 150$ eV, and so the gas within them is highly neutral. These systems, although rare, are extraordinarily rich in information. They have multiple velocity components, a wide range of metal lines (with a wide range of ionization states), and sometimes even molecular hydrogen.

DLAs are now understood to be absorption from the interstellar medium of galaxies. The lines therefore provide an alternate selection of galaxies: one that is weighted by *geometric cross-section* rather than stellar luminosity. They are therefore typically low-surface brightness galaxies with relatively low star formation rates, requiring exceptionally deep observations to identify their emission in conventional galaxy surveys. DLAs therefore provide an unbiased census of the neutral gas in the Universe. Moreover, based on the observed column density distribution of H I absorbers, most of the neutral hydrogen after reionization resides in DLAs. Interestingly, the fraction of gas that remains neutral appears to vary little with redshift from $z \sim 5$ to the present day, although of course that must change at higher redshifts when the IGM itself becomes predominantly neutral. For our purposes, DLAs are crucial as the primary reservoir of neutral gas after the end of reionization.

4.4.3 Fluctuations in the Ionizing Background

Because ionizing photons can only travel finite distances and are generated by discrete sources, one naturally expects fluctuations in the amplitude (and possibly shape) of the ionizing background. In practice, these are very small at $z < 5$, because λ is relatively large (see §4.4.2) and the ionizing sources are relatively common (particularly galaxies, provided that their escape fraction of ionizing photons is non-zero). But these fluctuations inevitably come to be important at higher redshifts (especially toward the epoch of reionization), and so it is important to understand their implications.

These fluctuations are sourced by both large-scale density fluctuations and stochastic variations in the number counts of the sources; the latter is most important when the number of sources within $\sim \lambda^3$ is small. A simple estimate of the effects of the density field is to compute the variance of the source population over one attenuation length, $\bar{b}\sigma(R = \lambda)$, where $\bar{b}$ is the average bias of the sources. At $z \sim 3$, taking $\bar{b} \sim 3$ and $\lambda \sim 300$ comoving Mpc yields fractional fluctuations of $\sim 2\%$, which is indeed close to more precise numerical estimates and is mostly negligible. However, at $z \sim 6$, taking the same average bias but $\lambda \sim 50$ Mpc implies fluctuations of $\sim 10\%$.

4.4.4 Helium-Ionizing Photons

About 7% by number of atoms or 24% by mass of the IGM gas is composed of helium atoms. Helium's first ionization potential is 24.6 eV and second is 54.4 eV. Photons above these threshold can therefore also interact with these species. The former is sufficiently close to the HI threshold that even stellar sources can ionize the first electron, provided that they can do the same to HI. However, normal stars do not produce significant numbers of photons above 54.4 eV to ionize HeII; that requires black holes.

The ionization cross-section for HeII follows the same form as in equation (4.16) (or the approximation in eq. 4.18). Like $\sigma_{\rm HI}$, this cross-section also scales as ν^{-3} near threshold. Note that for $\nu > \nu_{\rm HeII}$, $n_{\rm He}\sigma_{\rm HeII}/n_{\rm H}\sigma_{\rm HI} > 1$, so despite its lower abundance more high-energy photons ionize HeII than HI, even when their ionization fractions are equal.

HeII is also more difficult to keep ionized because it recombines faster than hydrogen; its case-B recombination coefficient is $\alpha_B = 1.53 \times 10^{-12}$ cm^3 s^{-1} at $T = 20,000$ K. The recombination timescale for gas at the mean density therefore remains smaller than the age of the Universe down to $z < 2$. Thus, HeII atoms may recombine many times over the age of the Universe.

Because of the large ionization cross section and rapid recombination time, the Universe remains optically thick to HeII-ionizing photons until relatively late (or, in other words, the mean-free-path of these photons is many times shorter than that of photons below the HeII ionization threshold). As a result, there is typically a substantial break in the ionizing background at the HeII ionization edge. Moreover, the HeII-ionizing background has much stronger fluctuations than the HI-ionizing background, both because of the short attenuation length and because only rare quasars contribute to it.

For the most part, the properties of this high-energy background have little effect on the HI Lyman-α forest; however, the photo-heating that occurs as the helium is ionized affects the hydrogen as well. The process is identical to that described above for HI reionization in equation (4.25) (except with $n_H \rightarrow n_{\rm He}$). Moreover, the hard spectra of quasars quite efficiently inject energy into the helium gas, so (despite the relative rarity of He atoms in the IGM) the total temperature increase can be comparable to that during hydrogen ionization. Once helium is reionized at $z \sim 3$, any influence of hydrogen reionization on the gas is largely erased.

4.5 METAL LINE SYSTEMS

So far we have focused exclusively on absorption by neutral hydrogen in the IGM. Can other elements be used to probe the IGM? Helium is an obvious candidate, but its Lyman-α line resides in the far-ultraviolet (with a rest wavelength of 304 Å) and is difficult to probe (although it has proved useful to study the HeII-ionizing background). These two are, of course, the primary elements produced in the Big Bang, but heavier elements do exist in the IGM owing to ejection and stripping from galaxies where they are produced through star formation.

The typical abundance of heavy elements in the IGM is small – with a median value $\langle Z \rangle \sim 10^{-3} Z_\odot$ – but the absorption is still substantial. If we make the simple assumption that the metals are uniformly distributed, we can repeat the fluctuating Gunn-Peterson approximation and find that the optical depth of an IGM patch to a given transition is

$$\tau_{X_i} = 0.097 f_i (1+\delta) \left(\frac{X}{3.6 \times 10^{-7}} \right) \left(\frac{f_{\rm osc}}{0.191} \right) \left(\frac{\lambda_i}{1548 \text{ Å}} \right) \left(\frac{1+z}{7} \right)^{3/2}, \tag{4.48}$$

where f_i is the fraction of the element in the appropriate ionization state, X is the abundance by number of the element relative to hydrogen, $f_{\rm osc}$ is the oscillator strength of the transition, and λ_i is its rest wavelength. The fiducial choices correspond to the stronger line in the CIV $\lambda 1548, 1551$ doublet with $Z = 10^{-3} Z_\odot$; Table 4.1 lists several other important transitions for low and high-z work. As we will see later, the assumption of a constant metallicity throughout the IGM is most certainly wrong, but it may be reasonable on the scale of a single absorbing system.

Clearly the optical depth can be substantial, even in relatively low density gas, provided that the gas is in the appropriate ionization state. In the diffuse IGM at low and moderate redshifts, these are highly-ionized states of the most common heavy elements, especially carbon, silicon and oxygen. C III and Si IV have ionization potentials of 47.888 and 45.142 eV, respectively; these two species should therefore evolve similarly, *unless* higher-energy photons are able to further ionize one but not the other. In fact, C IV and Si V require 64.492 and 166.77 eV to get ionized. The latter energy is relatively large, but once He II is ionized to He III the universe becomes transparent to photons that can ionize C IV which are still relatively common. We might therefore expect C IV and Si IV to be relatively abundant absorbers, at least until He II reionization completes at $z < 3$.

Both of these species are also very useful from an observational perspective, because they have doublet transitions redward of HI Lyman-α. A transition wth $\lambda_i > \lambda_\alpha$ is unaffected by HI absorption in the interval $z_{\rm min} < z < z_s$, where z_s is the redshift of the background source and

$$(1+z_{\rm min}) = (1+z_s) \frac{\lambda_\alpha}{\lambda_i}. \tag{4.49}$$

Absorbers in this range produce isolated absorption features against the red continuum of the source. Doublet transitions are particularly interesting because they make the species causing the absorption easy to identify, even in some cases without knowing anything about the HI absorption.

An exception to this rule is provided by oxygen, whose fifth ionization state is an important observational tracer despite the fact that its primary absorption feature is blueward of HI Lyman-α at $1032, 1038$ Å. This transition therefore suffers from contamination by the HI forest, but because it is a doublet it can sometimes still be measured. It is particularly useful for constraining the properties of hot gas, because the ionization potential of O V is 77.413 eV.

In a neutral gas, the relevant ions are different. For example, C I has an ionization potential of 11.26 eV, so provided that there is a source of UV photons – even if the ionizing photons are attenuated – carbon atmos will preferentially turn into C II

(which has an ionization potential of 24.383 eV and so cannot be ionized in gas that is optically thick to hydrogen-ionizing photons). Iron, another common heavy element, and silicon, occupy their first ionization state for similar reasons.

A particularly interesting case is oxygen, whose first ionization potential is 13.618 eV – nearly equal to HI. As a result, these two species should be locked in charge exchange equilibrium through the interaction

$$\mathrm{O}^0 + \mathrm{H}^+ \leftrightarrow \mathrm{O}^+ + \mathrm{H}^0 \tag{4.50}$$

whose equilibration timescale is $\sim 1/k_{\rm ce} n_{\rm HI} \sim 2 \times 10^5 x_{\rm HI}(1+\delta)(1+z/7)^3$ yr, much shorter than the Hubble time (where $k_{\rm ce}$ is the collisional rate coefficient). Thus dual observations of OI and HI provide an estimate of the metallicity (or, if that can be guessed, of the neutral fraction of HI) even when that line is highly saturated.

Note that all of these transitions relevant to neutral gas are singlets, so they are more difficult to identify than the C IV and Si IV lines in an ionized gas. This means that the transitions must be identified in combination with each other (or H I); with the complication that the different elements may have different abundances

At moderate redshifts, C IV absorbers are the most commonly studied (primarily because they are the easiest to find), with metal absorption visible in most individual systems with $N_{\rm HI} > 10^{15}$ cm^{-2} at a metallicity $Z \sim 10^{-2} Z_\odot$. They can also be detected statistically in much less dense systems, implying a median metallicity in forest absorbers of $Z \sim 10^{-3} Z_\odot$. Many other transitions are detectable in higher column density systems, especially in the DLA range (where the neutral gas makes transitions like CII and OI useful, although these systems usually have many different absorption components, some of which are also highly ionized); these are well-understood as being due to internal metal enrichment of galaxies. OVI has also received intense attention as a possible proxy for the hot, collisionally-ionized gas in galactic winds.

Despite the relative wealth of observations of metal absorption, the physics behind metals in the IGM remains mysterious. The forest absorbers themselves correspond to gas near or above the mean cosmic density, and such sheets and filaments only fill a relatively small fraction of the volume. Thus, observations currently require only $> 10\%$ of space to be enriched with metals. The key question is how and when did this enrichment occur: many models appeal to winds from the first galaxies but more powerful winds from star-forming galaxies at lower redshift are also a plausible explanation. More precise measurements of the spatial distribution of the metals (especially in comparison to samples of galaxies), their abundance patterns, and the evolution toward higher redshift, may help to constrain or eliminate some of these models.

4.6 THE LYMAN-α FOREST AT $Z > 5$

We now turn to the Lyman-α forest at very high redshifts, approaching the time of reionization and the first galaxies. As equation (4.35) shows, the absorption is quite thick by $z \sim 5.5$ when $\tau_{{\rm eff},\alpha} \sim 2.6$ with only $\sim 7\%$ of the light transmitted.

Past that point, the forest thickens even more rapidly, so that very little light is transmitted.

Of course, this low-level of transmission is not uniform across the entire spectrum due to the density fluctuations in the cosmic web. The small pockets of residual transmission correspond to underdense regions in the IGM. At $z \sim 5$–6, these pockets are sufficiently common that the forest can still be used to measure the properties of the IGM, and in particular the ionizing background – which appears to be only a factor of two or so smaller than at lower redshifts (with $\Gamma_{12} \sim 0.5$).

Unfortunately, beyond that point the Lyman-α forest itself becomes too thick to model robustly; in fact it is so thick that one can no longer pick out individual absorbers, and it is better to simply use the fluctuating Gunn-Peterson approximation. If one then has a model for the volume-weighted probability distribution of the IGM density $p(\delta)$, the effective optical depth is simply given by equation (4.29). The function $p(\delta)$ is easy to describe qualitatively: it must peak near $\delta \sim 0$, with a long tail toward high densities (describing collapsed structures) and another tail toward underdense voids which is truncated below a value $\delta = -1$ (corresponding to space with no matter). Equation (4.28) shows that, with $\Gamma_{12} \sim 0.5$ at $z \sim 6$, $\tau_\alpha \sim 26(1+\delta)^2$ (ignoring the weak temperature dependence), requiring $-1 < \delta < -0.8$ for substantial transmission. Thus, the crucial piece of the integral involves the far-end of the low-density tail (note that these voids are actually in the nonlinear regime), which is very difficult to model robustly without large numerical simulations. Even then, to measure the mean neutral fraction of the entire IGM one must extrapolate to significantly higher densities, which constitutes a highly uncertain operation. Conservatively, the observed transmission requires only a very small neutral fraction, $x_{\rm HI} < 10^{-4}$ at the mean density. Thus, the increasing optical depth of the forest with redshift is *not* necessarily a flag of the tail end of reionization; careful modeling of the forest is required to reach such a conclusion.

Table 4.1 Important IGM Metal-Line Transitions

Element	n_X/n_H ($\times 10^4$, for $Z_\odot$)	Ionization State	λ (Å)	$f_{\rm osc}$
Carbon	3.58	C II	1334.5	0.128
		C IV	1548.2*	0.191
		C IV	1550.8*	0.095
Oxygen	8.49	O I	1302.2	0.049
		O VI	1031.9*	0.133
		O VI	1037.6*	0.066
Silicon	0.33	Si II	1304.4	0.094
		Si IV	1393.8*	0.514
		Si IV	1402.8*	0.255
Iron	0.30	Fe II	1608.5	0.058
		Fe II	2344.2	0.114
		Fe II	2382.8	0.300

* Member of doublet

A few options can help to improve this measurement and extend the usefulness of the Lyman-α forest to higher redshifts. The first is to use a different aspect of the forest: one probe that appears promising is to use large-scale variations in the optical depth of the forest, which may be modulated by the contrast between neutral and ionized regions in the IGM. For example, some lines of sight at $z > 6$ show completely saturated absorption even in deep spectra, while others show clear transmission. Unfortunately, as described in §4.3.4, fluctuations in the absorption are dominated by the aliasing of small-scale modes in the density field, which tend to mask the underlying large-scale fluctuations. Moreover, the extremely underdense voids that allow transmission tend to lie in large-scale underdensities, which exaggerates their variance (i.e., they cluster just like rare, massive halos). Thus it is so far difficult to use these variations to constrain the neutral fraction.

A second option is to use a higher Lyman-series line: so far, Lyman-β (with $\lambda_\beta = 1026$ Å and $\tau_\beta/\tau_\alpha = 0.16$ at a fixed density) and Lyman-γ (with $\lambda_\gamma = 972$ Å and $\tau_\gamma/\tau_\alpha = 0.0558$ at a fixed density) have been used. With their smaller oscillator strengths, these lines can have considerably more transmission and so sample gas closer to the mean density; however, the primary difficulty is that they are visible only at $\lambda_{\rm obs} < \lambda_{\beta,\gamma}(1 + z_s)$, which is inside the Lyman-α forest of the same source (albeit at $z < \lambda_{\beta,\gamma}/\lambda_\alpha(1 + z_s)$, where the transmission is larger). One must therefore account for this unknown foreground absorption, which does introduce extra errors. Nevertheless, the higher Lyman-series lines do appear to be more sensitive than Lyman-α, and they indicate a steepening in the effective absorption of the IGM and hence possible stronger evidence for an increasing neutral fraction at $z > 6$.

One complication regarding these lines is that, because they probe slightly different densities than Lyman-α, they may also sample different temperatures if the gas is no longer isothermal ($\gamma \neq 1$). Indeed, as discussed in §4.3.1, the IGM is expected to have such a density-temperature relation once the gas relaxes after being heated during reionization. Because the temperature also affects the optical depth, this makes inferences about Γ more difficult (see equation 4.28). On the other hand, it also offers a route to *measure* this temperature-density relation and constrain the time of reionization that way (with the complication that denser regions may have reionized earlier than underdense regions).

Finally, instead of choosing weaker Lyman-series lines one can study rarer elements – the metal lines. With the forest saturated, it is no longer possible to associate these lines with HI features; however, they can still be detected individually as long as they appear redward of $\lambda_\alpha(1 + z_s)$ (see equation 4.49). Of course, one must then determine which species causes the observed line, e.g. by detecting multiple absorbers from the same redshift. Although this wavelength range pushes into near-infrared wavelengths for $z > 6$, searches for both OI lines and CIV doublets have been conducted. The results are intriguing but, so far, difficult to interpret: from $z \sim 6$ to $z \sim 5$, there is a rapid decrease in the density of OI lines, at least along some lines of sight, and a rapid increase in the density of CIV lines. Whether this represents evolution in the enrichment of the IGM, the ionizing background, or something else, remains unknown.

Two other probes of the ionization state of the IGM are useful as more direct

measurements of the reionization process: the so-called **red damping wing** (which refers to the Lyman-α absorption profile far to the red of line center, where the optical depth is of order unity even in a completely neutral medium) and the **proximity effect** (which refers to the highly-ionized zone surrounding individual bright sources). We discuss these probes in chapter 10.

Chapter Five

Primordial Stars

The formation of the first stars hundreds of millions of years after the Big Bang marks a crucial transition in the early Universe. Before this point, the Universe was elegantly described by a small number of parameters. But as soon as the first stars formed, complex chemical and radiative processes entered the scene. Today, 13.7 billion years later, we find very complex structures around us. Even though the present conditions in galaxies are a direct consequence of the simple initial conditions, the relationship between them was irreversibly blurred by complex processes over many decades of scales that cannot be fully simulated with present-day computers. Complexity reached its peak with the emergence of biology out of astrophysics.

The development of large scale cosmic structures occurs in three stages, as originally recognized by the Soviet physicist Yakov Zel'dovich. First, a region collapses along one axis, making a two-dimensional sheet. Then the sheet collapses along the second axis, making a one-dimensional filament. Finally, the filament collapses along the third axis into a virialized halo. A snapshot of the distribution of dark matter at a given cosmic time should show a mix of these geometries in different regions that reached different evolutionary stages (owing to their different densities). The sheets define the boundary of voids from where their material was assembled; the intersection of sheets define filaments, and the intersection of filaments define halos – into which the material is ultimately drained. The resulting network of structures, shown in Figure 4.1, delineates the so-called "cosmic web." Gas tends to follow the dark matter except within shallow potential wells into which it does not assemble, owing to its finite pressure. Computer simulations have provided highly accurate maps of how the dark matter is expected to be distributed since its dynamics is dictated only by gravity, but unfortunately, this matter is invisible. As soon as ordinary matter is added, complexity arises because of its cooling, chemistry, and fragmentation into stars and black holes. Although theorists have a difficult time modeling the dynamics of visible matter reliably, observers can monitor its distribution through telescopes. The art of cosmological studies of galaxies involves a delicate dance between what we observe but do not fully understand and what we fully understand but cannot observe. The next several chapters will describe this methodology.

When a dark matter halo collapses, the associated gas falls in at a speed comparable to V_c in equation (3.32). When multiple gas streams collide and settle to a static configuration, the gas shocks to the virial temperature $T_{\rm vir}$ in equation (3.33) – at which it is supported against gravity by its thermal pressure. At this temperature, the Jeans mass equals the total mass of the galaxy. In order for fragmentation

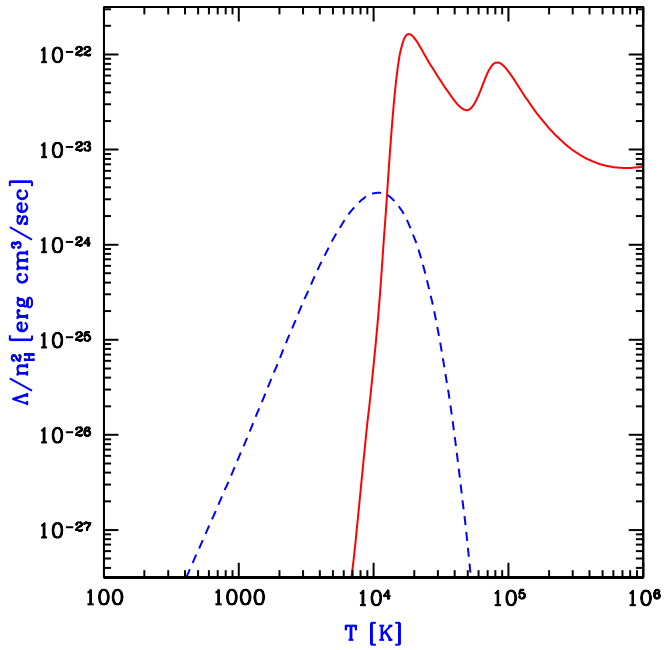

Figure 5.1 Cooling rates as a function of temperature for a primordial gas composed of atomic hydrogen and helium, as well as molecular hydrogen, in the absence of any external radiation. We assume a hydrogen number density $n_H = 0.045\ \mathrm{cm}^{-3}$, corresponding to the mean density of virialized halos at $z = 10$. The plotted quantity Λ/n_H^2 is roughly independent of density (unless $n_H > 10\ \mathrm{cm}^{-3}$), where Λ is the volume cooling rate (in erg/sec/cm^3). The solid line shows the cooling curve for an atomic gas, with the characteristic peaks due to collisional excitation of hydrogen and helium. The dashed line shows the additional contribution of molecular cooling, assuming a molecular abundance equal to 1% of n_H.

to occur and stars to form, the collapsed gas has to cool and get denser until its Jeans mass drops to the mass scale of individual stars.

Cooling of the gas in the Milky Way galaxy (the so-called "interstellar medium") is controlled by abundant heavy elements, such as carbon, oxygen, or nitrogen, which were produced in the interiors of stars. However, before the first stars formed there were no such heavy elements around and the gas was able to cool only through radiative transitions of atomic and molecular hydrogen. Figure 5.1 illustrates the cooling rate of the primordial gas as a function of its temperature. Below a temperature of $\sim 10^4$K, atomic transitions are not effective because collisions among the atoms do not carry sufficient energy to excite the atoms and cause them to emit radiation through the decay of the excited states. Since the first gas clouds around the Jeans mass had a virial temperature well below 10^4K, cooling and fragmentation of the gas had to rely on an alternative coolant with sufficiently low energy levels

and a correspondingly low excitation temperature, namely molecular hydrogen, H_2. Hydrogen molecules could have formed through a rare chemical reaction involving the negative hydrogen (H^-) ion in which free electrons (e^-) act as catalysts. After cosmological recombination, the H_2 abundance was negligible. However, inside the first gas clouds, there was a sufficient abundance of free electrons to catalyze H_2 and cool the gas to temperatures as low as hundreds of degrees K (similar to the temperature range presently on Earth).

The hydrogen molecule is fragile and can easily be broken by UV photons (with energies in the range of 11.26-13.6 eV), to which the cosmic gas is transparent even before it is ionized.[27] The first population of stars was therefore suicidal. As soon as the very early stars formed and produced a background of UV light, this background light dissociated molecular hydrogen and suppressed the prospects for the formation of similar stars inside distant halos with low virial temperatures $T_{\rm vir}$. However, illumination by X-rays could produce free electrons that would catalyse the formation of additional molecular hydrogen.

In order to understand how structures proceed from the first stars to subsequent generations, we must therefore understand *feedback processes* – in this case, UV and X-ray radiative feedback. We will therefore examine in some detail the growth of these radiation backgrounds and how they may affect star and galaxy formation. In particular, we will discuss how the chemistry of cooling changes dramatically when halos with $T_{\rm vir} > 10^4$K formed. In such objects, atomic hydrogen was able to cool the gas in them and allow fragmentation even in the absence of H_2 – such halos are thus immune to the radiation background.

The youngest stars in the Milky Way galaxy, with the highest abundance of elements heavier than helium (referred to by astronomers as 'metals') – like the Sun, were historically categorized as Population I stars. Older stars, with much lower metallicity, were called Population II stars, and the first metal-free stars are referred to as Population III.

Of course, because these same stars also produce heavy elements, which affect the chemistry and cooling of the gas, we must also track chemical feedback: how these elements were generated inside dark matter halos and how mechanical processes, most likely from supernovae or AGN, distributed these heavy elements within their parent halos and throughout the intergalactic medium (and hence the halos that assemble from it).

When these feedback mechanisms are included, the first structures to form stars likely cannot continue to do so, at least for a time: only later, possibly when atomic cooling becomes possible, will larger halos develop in which self-sustaining "galaxies" can form. These long-lived objects will be much easier to observe than their predecessors and hence provide an important marker in structure formation, especially for observers.

Unlike the previous chapters, in which much of the physics is clearly understood with reference to observations at low or moderate redshifts, the first stars and galaxies – and their immediate descendants – have yet to be observed. We will therefore focus in this section on the fundamental physical processes that shaped early star formation, but only sketch a preliminary picture of how these processes fit together in producing the first galaxies in the real Universe.

5.1 THE FIRST STARS: FROM VIRIALIZED HALOS TO PROTOSTARS

We have already seen that gravity drives the bottom-up hierarchy of structure formation characteristic of CDM cosmologies; however, at lower masses, gas pressure delays the collapse. The first objects to collapse are those just above the mass scale that allows cooling. Such objects reach virial temperatures of several hundred degrees and can fragment into stars only through cooling by molecular hydrogen. If this occurs faster than the dynamical time, the halo gas will collapse rapidly to form stars. In other words, there are two independent minimum mass thresholds for star formation: the filter mass (related to accretion and discussed in §3.2) and the cooling mass (related to the ability of the gas to cool over a dynamical time). For the very first objects, the cooling threshold is somewhat higher and sets a lower limit on the halo mass of $\sim 5 \times 10^4 M_\odot$ at $z \sim 20$. We will next examine this process in detail.

5.1.1 Chemistry of the Primordial Gas

The primordial gas out of which the first stars were made had 76% of its mass in hydrogen and 24% in helium and did not contain elements heavier than lithium. This is because during Big-Bang nucleosynthesis, the cosmic expansion rate was too fast to allow the synthesis of heavier elements through nuclear fusion reactions.

Before elements heavier than helium (denoted by astronomers as 'metals') were produced in stellar interiors, the primary molecule to reach sufficient abundance to affect the thermal state of the pristine cosmic gas was molecular hydrogen, H_2. The dominant H_2 formation process is

$$\mathrm{H} \quad + \quad \mathrm{e}^- \rightarrow \mathrm{H}^- \quad + \quad h\nu, \tag{5.1}$$

$$\mathrm{H}^- \quad + \quad \mathrm{H} \rightarrow \mathrm{H}_2 \quad + \quad \mathrm{e}^-, \tag{5.2}$$

where free electrons act as catalysts. We let the ionized fraction of hydrogen be $x_{\rm HII} = n_{\rm HII}/n$, where $n = n_{\rm HI} + n_{\rm HII}$ is the total abundance of hydrogen nuclei, and write the molecular fraction as $f_{\rm H_2} = n_{\rm H_2}/n$. Then, we can write our simplified reaction network as

$$\dot{x}_{\rm HII} = -\alpha_B n x_{\rm HII}^2 \tag{5.3}$$

$$\dot{f}_{\rm H_2} = \tilde{k} n (1 - x_{\rm HII} - 2 f_{\rm H_2}) x_{\rm HII}, \tag{5.4}$$

where the first equation follows recombinations (and hence the free electron fraction) and the second includes the steps of molecular hydrogen formation, which occurs at a net rate $\tilde{k}$. This net rate coefficient actually includes both equation (5.1), whose rate we shall call k_2 (for consistency with the literature) and equation (5.2), whose rate we shall call k_3. However, H^- is fragile and can be destroyed by CMB photons; we must therefore include a second channel in which the H^- does *not* lead to molecular hydrogen. This occurs at a rate

$$k_4 \approx 0.114 T_\gamma^{2.13} \exp(-8650\ {\rm K}/T_\gamma)\ {\rm s}^{-1}. \tag{5.5}$$

Thus, the net rate of H_2 formation is

$$\tilde{k} \approx k_2 \left[\frac{k_3}{k_3 + k_4/[(1 - x_{\rm HII})n]} \right], \tag{5.6}$$

where the second factor is the fraction of H^- that eventually forms H_2.

In reality, there are other channels to produce (and destroy) molecules. The set of important chemical reactions leading to the formation of H_2 is summarized in Table 5.1, along with the associated rate coefficients.[i] Detailed calculations require numerical integration of this network, but equations (5.3) and (5.4) provide some useful insight.

First, note that the ionized fraction is independent of $f_{\rm H_2}$, since the electrons only act as catalysts. Then, because $\dot{x}_{\rm HII} \propto x_{\rm HII}^2$, recombination will be very slow. This means that the reservoir of catalyzing electrons remains substantial for long periods of time (much larger than the recombination timescale): the solution with constant T and n (i.e., the inefficient cooling limit) is

$$x_{\rm HII}(t) = \frac{x_{\rm HII}^i}{1 + t/t_{\rm rec}^i}, \tag{5.7}$$

where $x_{\rm HII}^i$ is the initial ionized fraction (taken from cosmological calculations, as in Fig. 2.2) and

$$t_{\rm rec}^i = (x_{\rm HII}^i \alpha_B n)^{-1} \approx 2.2 \times 10^8 \left(\frac{1+z}{20}\right)^{-3} \left(\frac{\Delta}{200}\right)^{-1} \left(\frac{x_{\rm HII}^i}{2\times10^{-4}}\right)^{-1} \ {\rm yr} \tag{5.8}$$

is the recombination time at the initial ionized fraction. In the second part, we have assumed the gas sits at an overdensity of ~ 200, typical of virialized objects, used the residual ionized fraction following recombination (Fig. 2.2), and adopted a temperature, $T \approx 10^3$ K.

We can now substitute this expression into equation (5.4). The factor $(1 - x_{\rm HII} - 2f_{\rm H_2})$ remains near unity for the initial conditions and timescales of interest. Moreover, $\tilde{k}$ is roughly constant; in that case, the equation is integrable and yields

$$f_{\rm H_2} \approx f_{\rm H_2}^i + \frac{\tilde{k}}{\alpha_B} \ln(1 + t/t_{\rm rec}^i), \tag{5.9}$$

where $f_{\rm H_2}^i \approx$ is the molecular fraction when the cloud forms (typically the IGM value after recombination, $\sim 6 \times 10^{-7}$, provided that there is not yet a radiation background from luminous sources). The molecular fraction therefore increases linearly with time when $t/t_{\rm rec}^i \ll 1$, but it slows to logarithmic growth past that point: the transition occurs when the electrons are incorporated into hydrogen atoms, removing the population of catalysts and hence dramatically slowing down H_2 formation. It occurs at a critical molecular fraction

$$f_{\rm H_2,s} \equiv \frac{\tilde{k}}{\alpha_B} \approx 3.5 \times 10^{-4} (T/1000\,{\rm K})^{1.52}, \tag{5.10}$$

where the "s" indicates saturation (though in actuality $f_{\rm H_2}$ does continue to increase slowly). In practice, the nominal recombination time inside these objects is rather close to the Hubble time, so the electrons are used up rather quickly in the denser centers of the halos, where molecule formation is also fastest. Thus most virialized objects reach this "saturation" limit.

[i] Table 5.2 in §5.3.2 shows the same for deuterium mediated reactions. These should be included in detailed calculations but have only minor effects on the star formation picture described in this section.

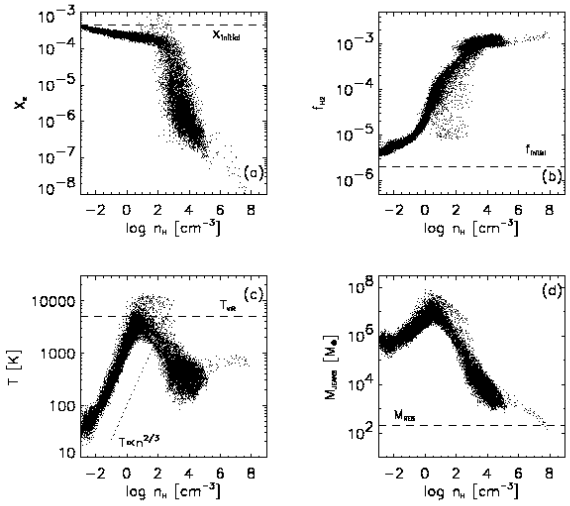

Figure 5.2 Gas properties during dense cloud collapse in a numerical simulation of first star formation. *(a)* Free electron abundance; note the rapid decline at $n \sim 10^3$ cm^{-3}, thanks to efficient recombinations. *(b)* Molecular fraction $f_{\rm H_2}$. The fraction increases rapidly during cloud collapse until the saturation value (eq. 5.10) is reached, when recombinations remove the free electron catalysts. *(c)* Gas temperature as a function of number density. Note the strong clump at $T \sim 500$ K and $n \sim n_{\rm cr}$, when radiative cooling becomes inefficient so the evolution stalls. *(d)* The Jeans mass for this gas; note that $M_J \approx 10^3\ M_\odot$ for gas in the aforementioned stalling stage. Figure credit: Bromm et al. 2002, ApJ, 564, 23.

The upper two panels of Figure 5.2 illustrate this process in a numerical simulation of the formation of the first stars. Panel *(a)* shows the free electron fraction in a collapsing gas cloud: it remains near the initial value (shown by the horizontal dashed line) for a while as the density increases before falling rapidly at $n > 10^3$ cm^{-3}, where recombinations become efficient. Panel *(b)* shows the molecular fraction, which increases steadily at low densities (and therefore early times in the collapse process) before reaching a limiting value near $f_{\rm H_2,s}$ in the densest part of the clump.

5.1.2 Cooling and Collapse of Primordial Gas

The next question is how much H_2 is required to allow the gas to cool and form stars. Cooling proceeds when an H_2 molecule is rotationally or vibrationally excited through a collision with another particle. If the subsequent de-excitation is radiative (and the cloud is optically thin), the cloud will lose energy and cool; if it is collisional, the cloud retains the energy, so no cooling occurs. In low den-

Table 5.1 Important reaction rates for Hydrogen species as functions of temperature T in K [with $T_\xi \equiv (T/10^\xi \mathrm{K})$]. For a comprehensive list of additional relevant reactions, see Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **467**, 522 (1996); Haiman, Z., Thoul, A. A., & Loeb, A., *Astrophys. J.* **464**, 523 (1996); and Abel, T. Anninos, P., Zhang, Y., & Norman, M. L. *Astrophys. J.* **508**, 518 (1997).

	Reaction	Rate Coefficient ($\mathrm{cm^3 s^{-1}}$)
(1)	$H + e^- \rightarrow H^+ + 2e^-$	$5.85 \times 10^{-11} T^{1/2} \exp(-157,809.1/T)(1+T_5^{1/2})^{-1}$
(2)	$H^+ + e^- \rightarrow H + h\nu$	$8.40 \times 10^{-11} T^{-1/2} T_3^{-0.2} (1+T_6^{0.7})^{-1}$
(3)	$H + e^- \rightarrow H^- + h\nu$	$1.65 \times 10^{-18} T_4^{0.76+0.15\log_{10} T_4 - 0.033 \log_{10}^2 T_4}$
(4)	$H + H^- \rightarrow H_2 + e^-$	1.30×10^{-9}
(5)	$H^- + H^+ \rightarrow 2H$	$7.00 \times 10^{-7} T^{-1/2}$
(6)	$H_2 + e^- \rightarrow H + H^-$	$2.70 \times 10^{-8} T^{-3/2} \exp(-43,000/T)$
(7)	$H_2 + H^+ \rightarrow H_2^+ + H$	$2.40 \times 10^{-9} \exp(-21,200/T)$
(8)	$H_2 + e^- \rightarrow 2H + e^-$	$4.38 \times 10^{-10} \exp(-102,000/T) T^{0.35}$
(9)	$H^- + e^- \rightarrow H + 2e^-$	$4.00 \times 10^{-12} T \exp(-8750/T)$
(10)	$H^- + H \rightarrow 2H + e^-$	$5.30 \times 10^{-20} T \exp(-8750/T)$

sity gas, collisions are sufficiently rare that the first channel dominates, and the cooling rate is proportional to n^2 because all of the molecules occupy low excitation states. Once collisions become important, the level populations shift to local thermodynamic equilibrium (LTE), and the cooling rate becomes proportional to n because the emergent intensity approaches the blackbody value. The transition occurs at the *critical density*, which is only a function of temperature; it corresponds to $n_{\rm cr} \approx 10^4$ cm^{-3} for the temperatures of interest to primordial star formation. Figure 5.3 shows how the cooling rates depend on density and temperature: note how the higher density rates approach the Local Thermodynamic Equilibrium (LTE) value near $n_{\rm cr}$. The initial stages of cloud formation therefore lie in the low-density regime where cooling is efficient.

A halo can collapse from the overdensities characteristic of virialization to those characteristic of stars only if cooling can occur much faster than the timescale over which the halo grows (and therefore accumulates more thermal energy). The latter is comparable to the Hubble time. The cooling time depends on the reaction networks discussed in the previous section. But the characteristic temperature to which H_2 radiation can drive gas is hundreds of K, because the two lowest rotational energy levels in H_2 have an energy spacing of $E/k_B \sim 512$ K. A reasonable approximation to the cooling time in a virialized halo is

$$t_{\rm cool} \approx 5 \times 10^4 f_{\rm H_2} \left(\frac{1+z}{20}\right)^3 \left(\frac{\Delta}{200}\right) \left(1 + \frac{10 T_3^{7/2}}{60 + T_3^4}\right)^{-1} \exp(512\,{\rm K}/T)\ {\rm yr}, \tag{5.11}$$

where $T_3 = T/(10^3\,{\rm K})$ and the temperature factors result from quantum mechanical calculations of the H_2 collisional excitation rates.

The relevant comparison to determine whether a gas cloud will collapse rapidly

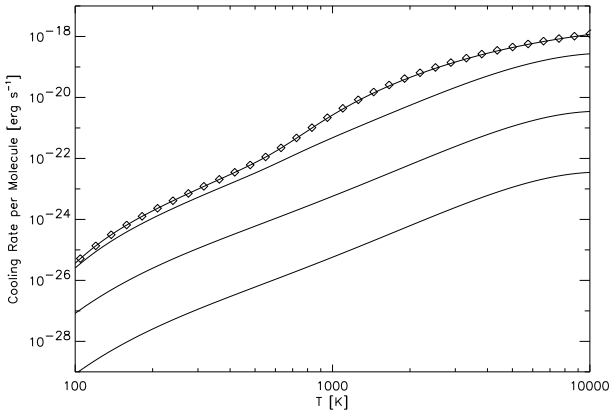

Figure 5.3 Cooling rate from H_2 per molecule. The solid lines show $n = 10^{-1}$, 10^1, 10^3, and 10^5 cm^{-3}, from bottom to top. The diamonds show the cooling rate in LTE; note how the cooling function approaches this limit when $n > n_{\rm cr}$ due to the transition to LTE. Figure credit: Bromm et al. 2002, ApJ, 564, 23.

to form stars is the dynamical time of the system, $t_{\rm dyn} \approx 1/\sqrt{G\rho}$ (with $\rho \sim m_H n$), which describes how rapidly gravity can adjust the configuration of the system. If $t_{\rm cool} > t_{\rm dyn}$, the cloud can adjust to the (slow) cooling quasi-statically. It will contract slowly, maintaining a constant Jeans mass, so that $T \propto \rho^{1/3}$. If, on the other hand, $t_{\rm cool} < t_{\rm dyn}$, the gas cloud will lose all its thermal energy much faster than gravity can adjust the configuration. As the pressure support vanishes, the cloud will collapse to much higher densities in roughly the free-fall time. We note as well that this argument is much broader than this particular application: it provides a useful minimal criterion for galaxy formation in a wide range of contexts.

In the present case, the relevant dynamical time is the Hubble time, t_H, because the cooling begins as soon as the cloud reaches high densities (or over a virialization time). Even after the halo forms, it will continue to accept gas (and thermal energy) and grow over roughly the same timescale. Using equation (5.11), the critical molecular fraction for rapid cooling to occur is

$$f_{{\rm H}_2,c} \approx 1.6 \times 10^{-4} \left(\frac{1+z}{20}\right)^{-3/2} \left(\frac{\Delta}{200}\right) \left(1 + \frac{10 T_3^{7/2}}{60 + T_3^4}\right)^{-1} \exp(512\ {\rm K}/T). \tag{5.12}$$

If a halo is able to form enough H_2 so that $f_{\rm H_2} > f_{{\rm H}_2,c}$, it will cool rapidly and form dense, highly molecular clouds. If not, it will remain a dense, virialized clump until it can surpass that threshold. We term such clumps *minihalos*.

Figure 5.4 shows that detailed numerical simulations of the early stages of structure formation confirm this picture. Each circle represents a single virialized object in the simulation; the filled circles contain dense clouds, while the open ones do

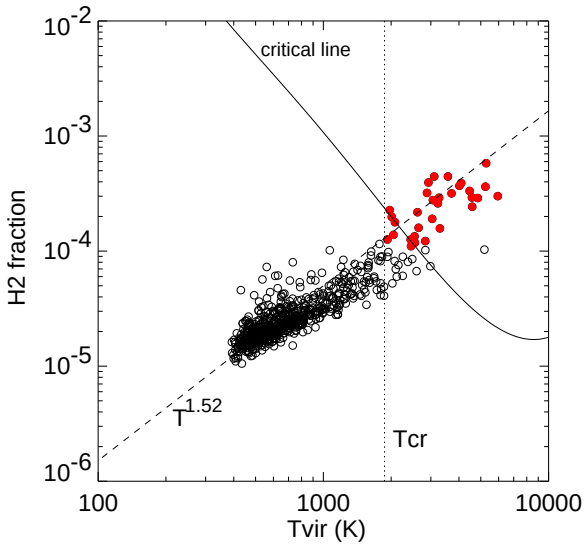

Figure 5.4 Molecular hydrogen fraction as a function of virial temperature for virialized halos inside a cosmological simulation at $z = 17$. The circles show results for individual halos; the filled circles contain dense (presumably star-forming) clouds, while the open circles do not. The dashed line shows the saturation limit $f_{\mathrm{H}_2,c}$ of eq. (5.10), while the solid line shows the critical molecular fraction for cooling to be rapid (see Eq. 5.12). The vertical dotted line show the critical virial temperature to host star-forming clouds. Figure credit: Yoshida et al. 2003, ApJ, 592, 645.

not. The dashed line shows the saturation limit for the molecular fraction: clearly the simulated halos lie remarkably close to this estimate, with the scatter likely due to variations in the accretion history of halos. The solid line shows the critical cooling threshold required at each virial temperature. It provides a remarkably accurate criterion to determine which halos can host dense, star-forming clouds.

However, these simulations find that $< 30\%$ of halos lying above the threshold still do not host star-forming clouds (the open circles in the upper right of Fig. 5.4), while some that lie below the nominal curve do have such clouds. These can be understood in terms of the accretion histories of the halos: recall that the cooling must balance the thermal energy gained throughout (ongoing) halo growth. Those halos accreting gas very rapidly may not be able to form dense clouds even if they are massive.

5.1.3 The Collapse of Dense Clouds

Cloud collapse continues until cooling becomes inefficient and thermal pressure significant. The minimum temperature achievable by H_2 cooling is $T \sim 200$ K, because the energy spacing of the first two rotational levels of that molecule is ~ 512 K (the limit is somewhat smaller than that value because of the high-velocity tail of the Maxwell-Boltzmann distribution). The characteristic density when cooling becomes inefficient is the critical density $n_{\rm cr} \approx 10^4$ cm^{-3} defined in the previous section, where collisions become frequent enough to maintain local thermodynamic equilibrium. At yet higher densities, the radiative intensity must follow the blackbody law, so the cooling rate is only linearly proportional to density (see Fig. 5.3).

With the decrease in the cooling rate, the gas cloud stalls or "loiters" at or near $n_{\rm cr}$. This stage is illustrated in panel *(c)* of Figure 5.2, which shows a phase diagram of the gas in a numerical simulation of these stages in the formation of the first stars. In the early stages (i.e., gas at low density in this diagram), cooling is inefficient (with a rate proportional to n^2), so the temperature roughly obeys the adiabatic relation $T \propto n^{2/3}$ (shown by the dotted line here). Once the density increases enough for H_2 cooling to become efficient, the temperature falls to $T \sim 200$ K, where it stalls as LTE is reached near the critical density.

Further collapse requires enough mass to accumulate for gravity to overcome the roughly constant pressure of this growing clump – in other words, until the mass of the clump exceeds the local Jeans mass, $M_J \approx c_s t_{\rm coll}$ (see §3.2). For gas in this clump, that is

$$M_J \approx 700 \left(\frac{T}{200\ {\rm K}}\right)^{3/2} \left(\frac{n}{10^4\ {\rm cm}^{-3}}\right)^{-1/2} M_\odot. \qquad (5.13)$$

Once the clump grows beyond this point, gravity drives further, rapid collapse on the dynamical timescale $t_{\rm coll}$ of the cloud.

To this point in the collapse, "first star formation" poses a physics problem with well specified initial conditions that can be solved on a computer. Starting with a simulation box in which primordial density fluctuations are realized (based on the initial power spectrum of density perturbations), one can reliably simulate the col-

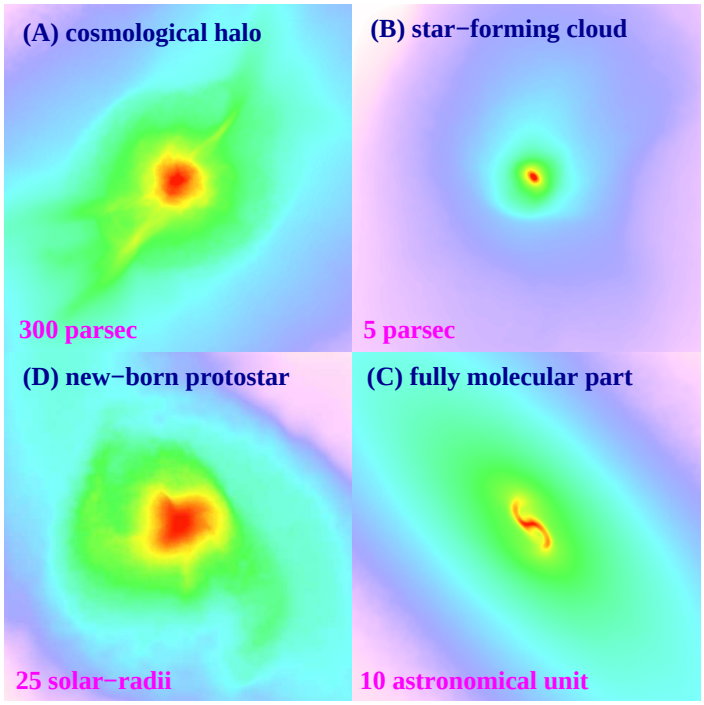

Figure 5.5 Projected gas distribution around a primordial protostar from a numerical simulation. Shown is the gas density (shaded so that dark grey denotes the highest density) of a single object on different spatial scales: *(a)* the large-scale gas distribution around the cosmological mini-halo; *(b)* the self-gravitating, star-forming cloud; *(c)* the central part of the fully molecular core; and *(d)* the final protostar. Figure credit: Yoshida, N., Omukai, K., & Hernquist, L. *Science* **321**, 669 (2008).

lapse by including the chemistry, gravitational dynamics, and thermodynamics of the gas. The top two panels in Figure 5.5 show these stages of collapse in a typical cosmological minihalo with $\sim 10^6 M_\odot$ in such a numerical simulation. Generically, the collapsing region makes a central massive clump with a typical mass of hundreds of solar masses, where the clump lingers because its H_2 cooling time is longer than its collapse time.

5.2 THE FIRST STARS: FROM PROTOSTARS TO STARS

Although the journey that led to humanity's existence was long and complicated, one fact is clear: our origins are traced to the production of the first heavy elements in the interiors of the first stars. Their formation is therefore a crucial milestone

in the Universe's history. The last section has put us on the cusp of understanding these objects – but, unfortunately, the evolution from that point is much more difficult to understand and still has many uncertainties.

Numerical simulations show that the protostellar core, with $T \sim 200$ K, gradually contracts at roughly constant temperature (owing to H_2 cooling) until $n > 10^8$ cm^{-3}. When the density becomes large enough for three-body processes to form H_2 through the reactions,

$$H + H + H \rightarrow H_2 + H, \tag{5.14}$$

$$H + H + H_2 \rightarrow H_2 + H_2. \tag{5.15}$$

The rate for the first of these reactions is $k_{3b} = 5.5 \times 10^{-29}$ cm^6 s^{-1}; the second is 1/8 as large. The timescale for this reaction equals the free-fall time at a critical density

$$n_{c,3} \approx \left(\frac{f_{H_2}^2 G m_p}{4 k_{3b}^2} \right)^{1/3}, \tag{5.16}$$

which is $\sim 10^8$ cm^{-3} for $f_{H_2} \sim f_{H_2,s}$. The molecular fraction then increases rapidly until it is near unity by the time $n \sim 10^{12}$ cm^{-3}, which one can estimate by setting $f_{H_2} \sim 0.5$ in equation (5.16). This stage is also shown in panel *c* in Fig. 5.5.

At this point, the large molecular fraction rapidly increases the cooling rate, allowing dynamical collapse. Numerical simulations show that a hydrostatic core of mass $< 10^{-2} M_\odot$ forms when the gas becomes optically thick to its own cooling radiation. This core forms the seed for a Population III star, but its subsequent evolution has proven much more difficult to track in numerical simulations. Not only is the dynamical time within the core very short, but the radiative feedback from the protostar couples to the gas, making the cooling processes more complex. Thus, the final products of even the well-posed problem of Population III star formation still have a fair amount of uncertainty. Here we will content ourselves with identifying the key issues in these final stages of formation.

5.2.1 A Single Protostar: No Feedback

We begin by considering the simplest case, in which the clump is assumed to form a single protostar. Theorists have made a good deal of progress in understanding how such a protostar would grow using a combination of numeric and analytic tools.

Star formation typically proceeds from the inside out, through the accretion of gas onto a central hydrostatic core. Whereas the initial mass of the hydrostatic core is very similar for primordial and present-day star formation, the accretion process – ultimately responsible for setting the final stellar mass – is expected to be rather different. It is common to parameterize the accretion rate as

$$\dot{m}_\star = \phi_\star \frac{m_\star}{t_{\rm ff}}, \tag{5.17}$$

where $\phi_\star$ is a parameter that depends upon the properties of the medium and $m_\star$ is the mass of the protostar. For a self-gravitating clump, the mass $m_\star \sim M_J \sim c_s^3 / \sqrt{G^3 \rho}$, the Jeans mass, so

$$\dot{m}_\star \sim c_s^3 / G \propto T^{3/2}. \tag{5.18}$$

A simple comparison of the temperatures in present-day star forming regions, in which heavy elements cool the gas to a temperature as low as $T \sim 10$ K, with those in primordial clouds ($T \sim 200 - 300$ K) already indicates a difference in the accretion rate of more than two orders of magnitude. This suggests that the first stars were probably much more massive than their present-day analogs. The key questions are to determine the accretion rate itself and estimate the duration over which it persists before radiative feedback from the central protostar (or star) shuts it off.

In order to estimate the accretion rate quantitatively, we need to determine ξ. The simplest solution is spherically symmetric accretion in a uniform medium onto a point mass, so-called *Bondi accretion*. A simple way to estimate how the accretion rate scales is to note that the protostar's gravity will overcome the pressure of the medium if the free-fall time $t_{\rm ff} \sim 1/\sqrt{G\rho}$ is smaller than the sound-crossing time $t_{\rm sc} \sim r/c_s$. This condition implies that infall will occur within a radius

$$R_{\rm acc} \sim \frac{Gm_\star}{c_s^2}. \tag{5.19}$$

The accretion rate will then be the surface area of a sphere at this radius, times the density of the medium, times the infall speed, which will be of order the sound speed. Thus

$$\dot{m}_\star \sim \frac{G^2 M^2 \rho}{c_s^3}. \tag{5.20}$$

We therefore have $\xi \sim (m_\star/\rho r^3)(t_{\rm sc}/t_{\rm ff})^3 \sim 1$, as expected.

Population III star formation is of course considerably more complicated than this simplest limit, as collapse proceeds in a virialized clump and is regulated by H_2 cooling. Nevertheless, it is possible to estimate the rate of collapse by using the numerical simulations to calibrate the models. We take a *self-similar* solution, in which all relevant physical quantities are power laws, because there is no characteristic length scale in the problem. We assume that the density field follows $\rho \propto r^{-k_\rho}$ and that the pressure follows $p \propto r^{-k_p}$. It follows that the solution is a polytrope, with $p \propto \rho_p^\gamma$.

The simulations show that the accretion process occurs subsonically and nearly isentropically, with an adiabatic index $\gamma \approx 1.1$ set by H_2 cooling. In hydrostatic equilibrium, the configuration therefore assumes a polytropic solution, with $P(r) \approx K\rho(r)^{1.1}$, and we immediately see that $\gamma_p = 1.1$ as well. Moreover, hydrostatic equilibrium

$$\frac{1}{\rho}\frac{dp}{dr} = -\frac{Gm_\star}{r^2}, \tag{5.21}$$

demands that $k_\rho = 2/(2-\gamma_p) \approx 20/9$ (i.e. the density structure is fairly close to an isothermal sphere) and $k_p = \gamma_p k_\rho$.

The constant K is set by the thermodynamics of the dense cloud during its "loitering" phase, with a fiducial value $K = 1.88 \times 10^{12} K_{\rm fid}$ in cgs units, where

$$K_{\rm fid} = \left(\frac{T}{300\ {\rm K}}\right)\left(\frac{10^4\ {\rm cm}^{-3}}{n}\right)^{0.1}. \tag{5.22}$$

This initial entropy, together with the initial density profile, ultimately determines the accretion rate onto the protostar. The hydrostatic equilibrium condition also requires that

$$\rho = \left[\frac{(3-k_\rho)k_p^3 K^3}{4\pi G^3 m_\star^2}\right]^{1/(4-3k_\rho)}, \tag{5.23}$$

Substituting into equation (5.17), we have

$$\dot{m}_\star = \frac{8\phi_\star}{\sqrt{3}}\left[\frac{(3-k_\rho)k_p^3 K^3}{2(2\pi)^{5-3\gamma_p}G^{3\gamma_p-1}}\right]^{1/2(4-3\gamma_p)} M^\xi, \tag{5.24}$$

$$\approx 0.026 K_{\rm fid}^{15/7}\left(\frac{m_\star}{M_\odot}\right)^{-3/7} M_\odot \ {\rm yr}^{-1}, \tag{5.25}$$

where in the second line we have used $\gamma_p = 1.1$ and evaluated $\phi_\star$ using the closest known self-similar solution to the early stages of accretion in simulations. Note, however, that the mass dependence is actually very sensitive to γ_p, varying from -0.37 to -0.49 for $\gamma_p = 1.09$–1.1. Nevertheless the solution clearly shows an important fact – and a key difference from low-mass star formation – that the accretion rate actually tapers off with time. The time required to build up a given stellar mass is

$$t \approx 27 K_{\rm fid}^{-15/7}\left(\frac{m_\star}{M_\odot}\right)^{10/7} {\rm yr}, \tag{5.26}$$

which matches detailed numerical simulations to within a factor of two or so in the early stages of protostar formation. Given that very massive Population III stars live for only a few Myr, this provides a *maximal* upper limit to the mass of the final star of $\sim 10^3\, M_\odot$, the accumulated mass over that lifetime, which depends on both the main sequence lifetime and the initial entropy of the gas.

In detail, provided that the core has some initial rotation, the gas falls onto an accretion disk rather than the star itself, and the resulting geometry may drive winds or other outflows, so the accretion rate estimated above is only accurate to a factor of order unity.

5.2.2 A Single Protostar: Radiative Feedback

The maximal mass estimate given above assumes that the protostellar (and stellar) radiation field does not affect the accretion. In the presence of this feedback, *can a Population III star ever reach this asymptotic mass limit?* The answer to this question is not yet known with any certainty, and it depends on how accretion is eventually curtailed by feedback from the star.

Before the onset of hydrogen fusion, the protostar must radiate away the gravitational energy accumulated by accretion, $L_{\rm acc} \approx G m_\star \dot{m}_\star / R_\star$, where $R_\star$ is the radius of the protostar. The outward radiation pressure on the gas can itself halt accretion if it balances the inward gravitational force. This is the *Eddington luminosity* L_E, representing the maximal luminosity of an accreting object. Assuming

for simplicity a fully ionized medium, force balance requires

$$\frac{Gm_\star m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{5.27}$$

where $\sigma_T = 0.677 \times 10^{-24}$ cm^2 is the Thomson cross-section for scattering a photon off an electron. Setting $L_{\rm acc} \approx L_E$ yields a critical accretion rate,

$$\dot{m}_{\star,E} \approx \frac{L_E R_\star}{Gm_\star} \sim 5 \times 10^{-3} \left(\frac{R_\star}{5R_\odot}\right) M_\odot \ {\rm yr}^{-1}, \tag{5.28}$$

where we have scaled $R_\star$ to a value typical of a very massive Population III star on the main sequence. Comparison of equations (5.28) and (5.25) suggests that radiative feedback can be crucial in halting accretion onto the protostar as it approaches the main sequence.

However, radiative feedback is likely to be unimportant at much earlier stages, because the protostar is much larger in size. For example, in the very early stages, when the opacity is dominated by H^- bound-free processes, the photosphere temperature is fixed at $T \sim 6000$ K because $\kappa_{H^-} \propto T^{14.5}$. Assuming that the protostar radiates as a blackbody, we then have

$$\frac{Gm_\star \dot{m}_\star}{R_\star} = 4\pi R_\star^2 \sigma_{\rm SB} T^4, \tag{5.29}$$

where $\sigma_{\rm SB}$ is the Stefan-Boltzmann constant. This yields $R_\star \approx 50(m_\star/M_\odot)^{1/3}\ R_\odot$ for $\dot{m}_\star \sim 0.005\ M_\odot$ yr^{-1}. Thus, we naively expect that radiative feedback will kick in only relatively late in the star formation process.

There are four distinct aspects of feedback exerted by a star on its gaseous environment:

- *Photodissociation of H_2:* As the protostar heats up it produces far-ultraviolet radiation that photodissociates H_2 (see §**??** below for a detailed discussion). Once molecular cooling turns off, the adiabatic index of the gas increases to $\gamma = 5/3$ (i.e., monatomic gas). This decreases the accretion rate (because the pressure increases more rapidly as the gas gets compressed), but numerical estimates and semi-analytic models show that the decline is rather modest. (This is not surprising given that the simple Bondi accretion problem described above also permits steady accretion when $\gamma = 5/3$.)

- *Lyman-α radiation pressure:* As we will discuss in detail in §10.1.1, the radiative transfer of Lyman-α photons is typically a very complex process when the optical depth is very large, as is true for a collapsing protostar surrounded by large quatities of neutral gas. The Lyman-α photons provide a substantial pressure, because they are trapped by the optically thick gas (and, on average, scattering off infalling gas blueshifts the photon, reducing the infall velocity of the gas). Indeed, they do not even escape by scattering through the gas column – rather, they escape when their frequency wanders so far from line center that the gas becomes effectively transparent. Because of these frequency shifts, the geometry of the flow plays an important role – as soon as a low-column density channel opens up in one direction, photons

can easily escape along that channel. Provided that accretion occurs through a disk, escape is most likely to occur along the polar direction, where the accretion rate is already quite small. Analytic estimates show that Lyman-α scattering can begin to slow the accretion when the core has $M \sim 20\ M_\odot$, but that the overall effect is small.

- *Ionization:* Once the protostar begins to produce ionizing photons, they will carve out an H II region in which the temperature is much larger than the surrounding neutral gas (typically $> 2 \times 10^4$ K; see §8.10 for a detailed discussion). This dramatically increases the pressure of the gas, which can cause the H II region to expand and drive off the gas that would otherwise accrete onto the protostar. The dynamics of the region depend upon the expansion velocity of the ionization front. If the front moves faster than about twice the ambient sound speed (of the neutral gas), then it has essentially no dynamical effect on the gas. This is known as an "R type" (or rarefied) front. Near a Population III protostar, the H II region begins in this regime, because it is expanding through gas falling in at the free-fall velocity $v_{\rm ff}$, which is highly supersonic.

  Eventually, the front reaches the radius where $v_{\rm ff} \sim 2c_s$, where the gas can respond to the ionization front, and a shock forms (this is a "D type," or dense front). Typically, the shock leads the ionization front, creating a dense shell of neutral gas into which the front propagates, with a bulk kinetic energy density comparable to the pressure inside the ionization front. A simple estimate for the point at which this shock halts accretion is thus when the thermal pressure gradient at the front exceeds the inward gravitational force. This is roughly the accretion radius $R_{\rm acc}$ defined in equation (5.19), with $T \sim 20,000$ K. Estimates of the ionizing luminosity of these protostars indicate that this limit is reached when $m_\star \sim 100\ M_\odot$.

  As before, the disk geometry of the accretion flow will play an important role in how this feedback mechanism plays out. The front will propagate fastest through the lowest column density of gas, which is along the polar axis, so accretion will first be suppressed there. In contrast, along the direction of the disk, the extreme column density of the disk "shadows" the flow, allowing accretion to continue. Provided that most of the accretion occurs through such a disk, the H II region will therefore not entirely halt the protostar's growth.

- *Photoevaporation of the Accretion Disk:* However, the same ionizing photons will heat the disk itself, evaporating gas from it and eventually shutting off accretion entirely. The rate at which this occurs depends upon the geometry of the disk and the spectrum of the protostar, but some calculations show that the disk evaporates when $m_\star \sim 150\ M_\odot$. As we will see below, this is very near the mass threshold for direct black hole formation when such stars die, so the details of the process may be very important.

Because these radiative feedback processes only become important late in the evolution of the first stars, they must generally be studied with simplified analytic

models rather than incorporated directly into ab initio simulations of Population III star formation. We therefore only have approximate estimates of their importance, and observations of these stars may be necessary to settle the physical uncertainties.

5.2.3 Multiple Protostars: Fragmentation

The models described above make one key assumption: that the collapsing material accretes onto a single object, the central protostar. However, we have argued that (with angular momentum) the accretion flow will generically organize itself into a disk. *Can this disk then fragment into multiple high-density clumps, or multiple protostars?* There are several possible mechanisms for fragmentation – gravitational instabilities, turbulence, and thermodynamic instabilities. All have now been implicated in numerical simulations showing fragmentation, but it is far from clear whether these are generic processes, and how severe the fragmentation is.

The classic way to gauge the importance of gravitational instability is the *Toomre criterion*. We sketch its significance here; more detailed derivations can be found in the references listed in the *Further Reading* section below. Consider a small patch inside a rotating gaseous disk. Let the patch have a radius r and mass $M = \pi\Sigma r^2$ (where $\Sigma = \rho/\Delta z$ is the surface density and Δz is the disk thickness). If we compress the patch by a factor δ, so $r \rightarrow r(1-\delta)$, the pressure increases by an amount

$$\Delta p \sim c_s^2 \delta\rho_0 \sim \delta c_s^2 \Sigma (\Delta z)^{-1}. \tag{5.30}$$

Thus, the excess pressure force per unit mass is

$$\frac{\nabla(\Delta p)}{\Sigma(\Delta z)^{-1}} \sim \frac{c_s^2\delta}{r}. \tag{5.31}$$

Meanwhile, the increase in the gravitational force per unit mass is $-GM\delta/r^2 \sim G\Sigma\delta$. Thus, the outward pressure counteracts gravity if

$$r < \frac{c_s^2}{G\Sigma} \equiv R_{\rm pr}. \tag{5.32}$$

This is just the classical Jeans analysis (§3.2) applied to a two-dimensional system: small wavelength modes are stabilized by pressure, while large wavelength modes are unstable to gravitational collapse.

However, in a rotating disk the angular momentum can stabilize these long wavelength modes. Assuming that our perturbation involved no external force (and hence torque), the internal spin angular momentum (generated by differential rotation across the patch) must be conserved. If Ω is the rotation speed, this is $J_s \sim \Omega r^2$.

As we compress the patch, conservation of angular momentum increases the rotation speed and thus creates a centrifugal barrier to further compression. To gauge how effective this is, we write the centripetal force per unit mass in terms of the conserved quantity J_s:

$$\frac{v^2}{r} \sim \frac{\Omega^2 r^2}{r} \sim \frac{J_s^2}{r^3}. \tag{5.33}$$

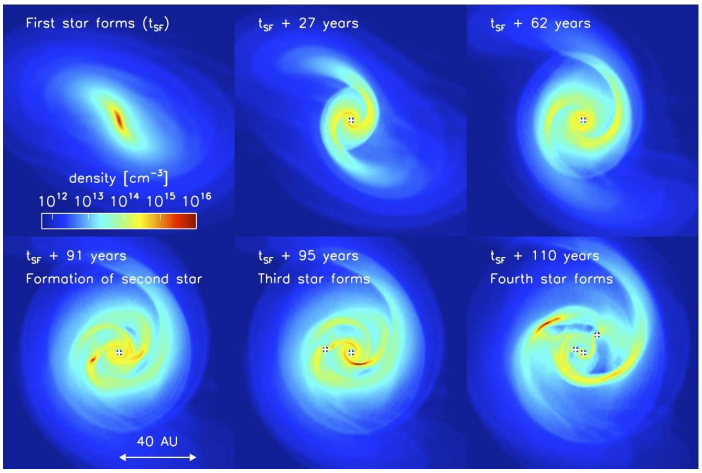

Figure 5.6 Density evolution in a 120 AU region around the first protostar in a numerical simulation of Population III star formation, showing the buildup of the protostellar disk and its eventual fragmentation at the times labeled in the diagram. Figure credit: Clark, P. C. et al., *Science* **331**, 1040 (2011).

Thus, the excess force as we compress the patch is $d(J_s^2/r^3)/dr \times \delta r$, which overcomes gravity and prevents further collapse if

$$r > \frac{G\Sigma_0}{\Omega^2} \equiv R_{\rm cen}. \tag{5.34}$$

We can only have an instability if $R_{\rm cen} > R_{\rm pr}$, so that a range of moderate wavelength perturbations cannot be stabilized neither by pressure or rotation. A more exact derivation shows that instability sets on if the *Toomre criterion*

$$Q \equiv \frac{c_s \kappa_e}{\pi G \Sigma} < 1 \tag{5.35}$$

Here κ_e is the epicycle frequency, or the rotation frequency for small perturbations around the equilibrium disk. For a Keplerian disk, $\kappa_e = \Omega = \sqrt{GM(r)/r}$, where $M((r)$ is the mass enclosed within a radius r.

Figure 5.6 shows this kind of gravitational fragmentation in a numerical simulation of the accretion disk around a Population III star. The disk very quickly exhibits spiral structure, common in self-gravitating disks, developing non-axisymmetric features and becoming locally unstable just ~ 100 yr after the formation of the first protostellar core and forming a second core separated by ~ 20 AU from the first. Figure 5.7 shows why: the top two panels show that the surface density and temperature of the disk remain roughly constant over time, except near its outskirts. This means the rate at which the disk can transport angular momentum (and hence material) inwards stalls, and the outer disk builds up more and more mass, quickly becoming gravitationally unstable ($Q \sim 1$ at $r \sim 20$ AU).

To continue fragmentation, the clump must still be able to rid itself of the thermal energy generated during collapse. At the characteristic densities of these disks

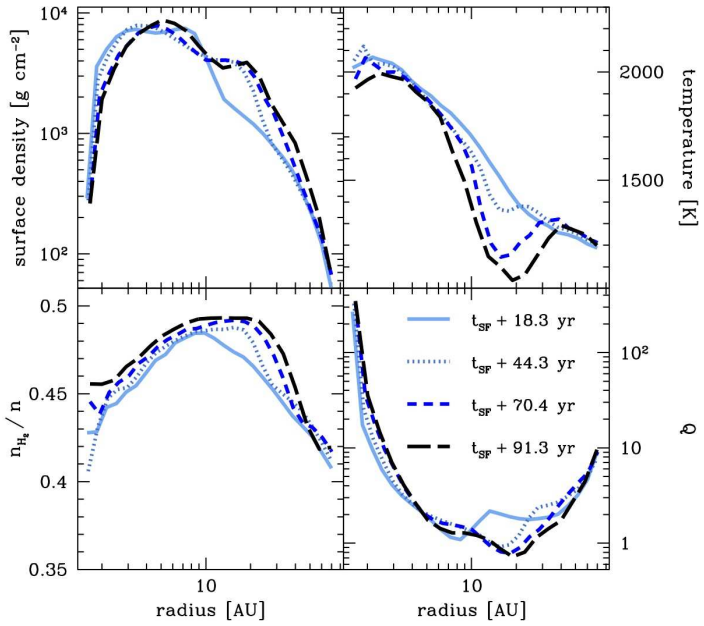

Figure 5.7 Radial profile of disk physical properties from the same simulation shown in Fig. 5.6, centered on the first protostellar core to form. Clockwise from upper left, the panels show the surface density, temperature, Toomre-Q parameter, and molecular fraction. Note how the disk parameters do not evolve strongly. The second core to form in the simulation forms within the region near $r \sim 20$ AU where $Q < 1$. Figure credit: Clark, P. C. et al., *Science* **331**, 1040 (2011).

($n \sim 10^{12}$–10^{14} cm^{-3}) a new cooling process dominates: *collision-induced emission* (CIE). This occurs when H_2 interacts with another species (H, He, or H_2) in a collision. The interacting pair briefly forms a "supermolecule" with a non-zero electric dipole, from which photons can be emitted or absorbed efficiently. Because the collision times are very short, the resulting radiation is emitted nearly in a continuum. This CIE radiation allows the gas to cool during the early stages of fragmentation, because the cooling time is substantially shorter than the dynamical time.

However, the continuum opacity of these same molecules prevents CIE cooling at $n > 10^{16}$ cm^{-3}. At this point, the gas does begin to heat up. However, at temperatures much above the $T \sim 10^3$ K characteristic of the disk (see the upper right panel in Fig. 5.7), H_2 begins to dissociate. Each such dissociation removes 4.48 eV from the gas, which keeps it near its original temperature because it is so highly molecular (see the lower left panel in Fig. 5.7).

Turbulence appears to be a third factor triggering instabilities and fragmentation. Such turbulence can be generated by "cold" accretion onto the host minihalo, where gas is funneled into the halo along filamentary channels and is not initially shock heated to the virial temperature of the halo. Instead, it collides with the central gas clump supersonically, triggering (typically subsonic) turbulent motions. Turbulence is known to be important in "normal" star formation at low redshifts leading to fragmentation of giant molecular clouds into protostellar cores with a wide range of initial masses. Some numerical simulations indicate that similar processes could cause fragmentation in the Population III regime.

5.2.4 The Initial Mass Function

Currently, we have no direct observational constraints on how the first stars formed at the end of the cosmic dark ages, in contrast to the wealth of observational data we have on star formation in the local Universe. Population I and II stars form out of cold, dense molecular gas that is structured in a complex, highly inhomogeneous way. The molecular clouds are supported against gravity by turbulent velocity fields and are pervaded by magnetic fields. Stars tend to form in clusters, ranging from a few hundred up to $\sim 10^6$ stars. It appears likely that the clustered nature of star formation leads to complicated dynamical interactions among the stars. The initial mass function (IMF) of Population I stars is observed to have a broken power-law form, originally identified by Ed Salpeter, with a number of stars $N_\star$ per logarithmic bin of star mass $m_\star$,

$$\frac{dN_\star}{d\log m_\star} \propto {m_\star}^{-\Gamma}. \tag{5.36}$$

Figure 5.8 shows some data in nearby star-forming regions, the only environment in which the IMF can be reliably measured, and the effective power-law index in these regions. The data are consistent with a broken power law

$$\Gamma \simeq \begin{cases} 1.35 & \text{for } m_\star > 0.5 M_\odot \\ 0.0 & \text{for } 0.008 M_\odot < m_\star < 0.5 M_\odot \end{cases}. \tag{5.37}$$

We shall take this as our fiducial model in the discussion, though we note that the form of the IMF at low masses is still unsettled. The lower cutoff in mass

corresponds roughly to the minimum fragment mass, set when the rate at which gravitational energy is released during the collapse exceeds the rate at which the gas can cool. Moreover, nuclear fusion reactions do not ignite in the cores of protostars below a mass of $\sim 0.08 M_{\odot}$, so-called "brown dwarfs". The most important feature of this IMF is that $\sim 0.5 M_{\odot}$ characterizes the mass scale of Population I and II star formation, in the sense that most of the stellar mass goes into stars with masses close to this value.

The ultimate goal of studies of the formation of Population III stars is to determine the analogous mass function for primordial stars. Unfortunately, we are far from converging on any robust predictions. Until recently, models of single protostar formation seemed to suggest that accretion would continue until $m_{\star} \sim 100\ M_{\odot}$, with the details determined by the initial entropy of the gas ($K_{\rm fid}$ in eq. 5.25) and by radiative feedback, with a plausible mass range from $\sim$ 20–$300\ M_{\odot}$. These masses – obviously much larger than the characteristic mass of present day stars – suggested that the first generation of stars to light up the Universe would be truly exotic objects.

However, the more recent studies of fragmenting disks suggest that the characteristic masses may be much smaller. Gravitational instability leads to several cores, each competing for the accreting gas. Turbulence may lead to an even wider range of initial protostar sizes. These cores themselves can interact, much as the stars in nearby open clusters do. In particular, three-body interactions tend to speed up smaller cores and move them into the outskirts of the core, where there is less gas to accrete. Meanwhile, the larger cores tend to sink to the center of the cloud, accreting more rapidly. This picture of "competitive accretion" may be important for high-mass star formation in the nearby Universe; if so, it may suggest that Population III star formation may also follow a power-law IMF with a broad range of stellar masses.

Nevertheless, it seems likely that the characteristic mass of high-redshift stars *must* be significantly larger than the present-day value of $\sim 0.5\ M_{\odot}$. The present-day value can be understood relatively easily as the minimum mass for collapse in the ~ 8 K molecular gas out of which these stars form (the minimum temperature is set by the cooling physics in molecular clouds). The Jeans mass provides a reasonable estimate of this value, but a more appropriate choice involves the *Bonnor-Ebert mass*,

$$M_{\rm BE} = 1.18 \frac{(kT/\mu m_p)^2}{p_0^{1/2} G^{3/2}}, \tag{5.38}$$

which is the largest mass that an isothermal gas sphere with a temperature T can have in hydrostatic equilibrium with an external gas pressure p_0. A Bonnor-Ebert sphere has a finite central density and size as it is confined by external pressure. Its maximum mass $M_{\rm BE}$ is 4.7 times smaller than the Jeans mass, but otherwise has the same scaling with density and temperature.

The temperature floor is expected to evolve with redshift, because radiative cooling cannot bring the temperature below the CMB temperature, to which all of the relevant lines couple. At $z = 30$, $T_{\rm CMB} = 82$ K, many times larger than the present day value (which note is well above the current CMB temperature). The

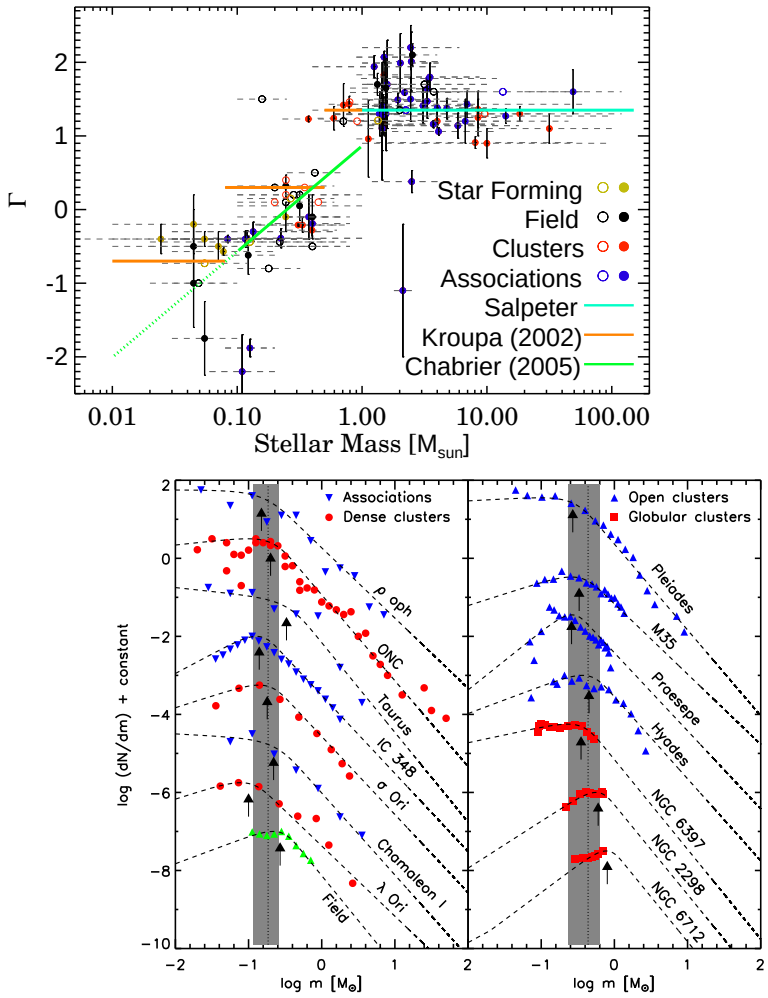

Figure 5.8 *Upper panel:* The derived power-law index, Γ, of the IMF in nearby star forming regions, clusters and associations of stars within the Milky Way galaxy, as a function of sampled stellar mass (points are placed in the center of log $m_\star$ range used to derive each index, with the dashed lines indicating the full range of masses sampled). The colored solid lines represent three analytical IMFs. *Bottom panel:* The present-day IMF in a sample of young star-forming regions, open clusters spanning a large age range, and old globular clusters. The dashed lines represent power-law fits to the data. The arrows show the characteristic mass of each fit, with the dotted line indicating the mean characteristic mass of the clusters in each panel, and the shaded region showing the standard deviation of the characteristic masses in that panel. The observations are consistent with a single underlying IMF. Figure credit: Bastian, N., Covey, K. R., & Meyer, M. R., *Ann. Rev. Astr. & Astrophys.* **48** (2010).

quantitative change in the Bonner-Ebert mass is not trivial to estimate, because it depends on the temperature-density relation in the clouds: for example, if the density structure is fixed, $M_{\rm BE} \propto T^{3/2}$, but if cooling proceeds isobarically, with $nT =$ constant, then $M_{\rm BE} \propto T^2$. This suggests that the characteristic fragmentation mass would increase to at least $\sim$ 16–50 $M_\odot$ at $z = 30$ (or even 10–20 $M_\odot$ at $z = 10$), well into the range of "high mass" stars by present-day standards.

5.3 THE SECOND GENERATION OF STARS: "POPULATION III.2"

The picture we have described so far assumes that the star formation process begins with the initial conditions characteristic of the high-redshift IGM: gas that is nearly neutral, with very little pre-existing H_2. These are, of course, the proper initial conditions for the first star-forming halos. But this picture depends rather sensitively on those assumptions, and it is likely that later generations of stars – still forming out of primordial gas – will form under different conditions.

The key is the initial ionization state of the gas. There are three important ways in which that can be much higher for these later stars. First, the first stars will produce a copious amount of ionizing photons, generating H II regions within and around their host dark matter halo. Any clumps that collapse within the ionized region will collapse from fully-ionized gas. Similarly, if these stars explode in supernovae, their powerful blastwaves will ionize the nearby gas (and possibly even trigger collapse). Finally, as larger mass halos form, star formation will shift to those larger objects. Above a virial temperature of $\sim 10^4$ K, the virialization shock itself will ionize the halo gas, again changing the initial conditions for cloud chemistry and collapse.

These initial conditions result in a different formation mode for primordial stars, often referred to as *Population III.2*, with a distinct initial mass function from the classic Population III.1 mode described earlier.

5.3.1 The Freeze-Out of Molecular Hydrogen

We showed in §5.1.1 that H_2 formation is catalyzed by the presence of free electrons. Thus, in gas that cools from a fully ionized state, molecule formation can proceed rapidly – even though at the initially high temperatures such molecules are dissociated.

Figure 5.9 shows numerical models of idealized isobaric cooling in primordial gas initially at $T \sim 10^4$ K (and hence ionized). As the gas cools, H_2 begins to form through the usual free electron channel, until its abundance saturates at $f_{\rm H_2} \sim 2 \times 10^{-3}$, regardless of the initial conditions. This "freeze-out" process indicates that the molecular fraction saturates at a non-equilibrium value.

In particular, $f_{\rm H_2}$ can no longer evolve once the timescale for H_2 formation ($t_{\rm form}$) and dissociation ($t_{\rm diss}$) become longer than the cooling and recombination timescales in the system. As in §5.1.1, the formation time can be approximated by $t_{\rm form} = f_{\rm H_2}/\dot{f}_{\rm H_2} \approx f_{\rm H_2}/(x_{\rm HII}\tilde{k}n)$. The dominant H_2 dissociation process is reaction (7) in Table 5.1, whose rate we will denote by k_7. Then $t_{\rm diss} = (k_7 x_{\rm HII} n)^{-1}$.

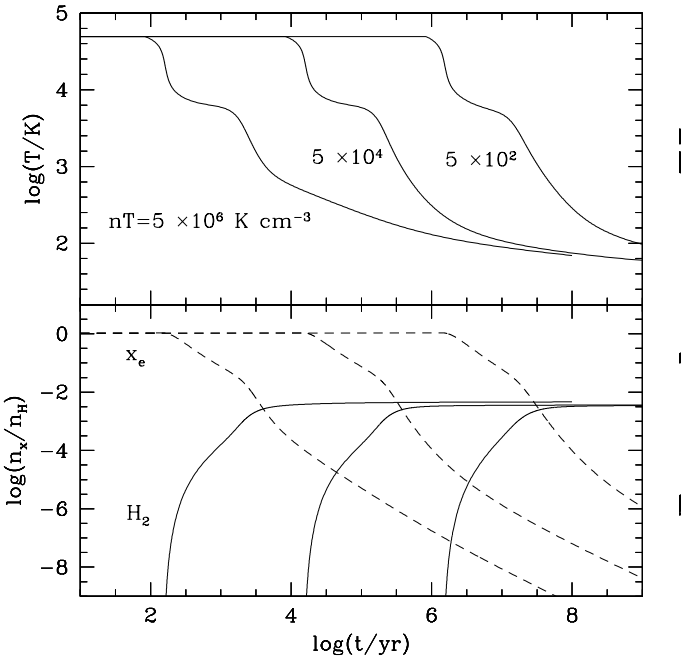

Figure 5.9 H_2 formation in initially ionized gas. The top panel shows the temperature evolution of gas at three different initial densities, assuming isobaric cooling (the three models are offset in time for clarity of presentation). The bottom panel shows the molecular fraction (solid curves) and the free electron fraction (dashed curves) for the same three models. Note how $f_{\rm H_2}$ approaches a constant limit in all three cases. Figure credit: Oh, S. P. & Haiman, Z., *ApJ*, **569**, 558 (2002).

The rate $t_{\rm diss}^{-1}$ decreases exponentially as the temperature drops, while the rates for cooling, recombination, and formation decrease only as power laws. This steep temperature dependence means that $t_{\rm diss}$ very suddenly becomes longer than $t_{\rm rec}$ and $t_{\rm cool}$ as the gas cools; the reaction rates demand that the resulting temperature is $T_{\rm freeze} = 3700\,{\rm K}$. Up to this point, the H_2 abundance remains in equilibrium, and the ratio of the reaction rates yields the value

$$f_{\rm H_2,freeze} \approx \frac{\tilde{k}(T_{\rm freeze}}{k_7(T_{\rm freeze})} \approx 2 \times 10^{-3}. \tag{5.39}$$

At lower temperatures, we know now that molecular hydrogen can no longer be destroyed. Since equilibrium demands that the formation and dissociation timescales would be comparable, this further implies that H_2 formation will also cease so long as its rate increases less slowly with temperature than cooling and recombination, which is readily shown by comparison of the reaction rates in Table 5.1. Thus, when $T < T_{\rm freeze}$, the molecular hydrogen abundance remains fixed at its (non-equilibrium) freeze-out value $f_{\rm H_2,freeze}$.

5.3.2 Deuterium and Cooling

The relatively high abundance of molecular gas already suggests that these pre-ionized systems can also eventually cool and form stars. However, there is an additional wrinkle that becomes important in these systems: deuterium. Unlike H_2, which is a symmetric molecule, HD has a permanent dipole moment which allows strong dipole rotational transitions with $\Delta J = \pm 1$, of lower energy than the $\Delta J = \pm 2$ quadrupole transitions of H_2 (the larger reduced mass of HD lowers this energy even further). The $J = 1 \rightarrow 0$ transition has an equivalent temperature of $\sim 130\,{\rm K}$, about four times smaller than the lowest energy transition of H_2. Thus, in principle, HD cooling can lower the temperature and hence mass scale of star formation substantially (recall that $M_J \propto T^{3/2}$ at fixed density, Eq. 5.13).

The most efficient method for HD to form is via the reaction

$$H_2 + D^+ \rightarrow HD + H^+. \tag{5.40}$$

This of course requires the simultaneous presence of molecular hydrogen and ionized deuterium. In the standard picture, which occurs entirely at low temperatures, the latter is very rare, and very little HD forms. However, in the present case, where all the deuterium begins ionized, the abundance of D^+ remains relatively large until very low temperatures. Thus, a substantial abundance of HD can build up, as illustrated in Figure 5.10. Table 5.2 provides reaction rates for the most important deuterium reactions.

Moreover, HD has several advantages as a coolant over H_2. First, it has a higher critical density, $n_{\rm crit,HD} \sim 10^6\,{\rm cm}^{-3}$, so rapid cooling continues to higher densities. Second, its dipole transitions are much more rapid, with a spontaneous decay rate $A_{10} \approx 5 \times 10^{-8}\,{\rm s}^{-1}$. This allows rapid cooling even at low abundances: at the levels shown in Figure 5.10, the gas can easily cool to the CMB temperature over a relatively short time. To see this, let us assume for simplicity that the gas, at temperature T, is in LTE, so that the level populations in the ground and first

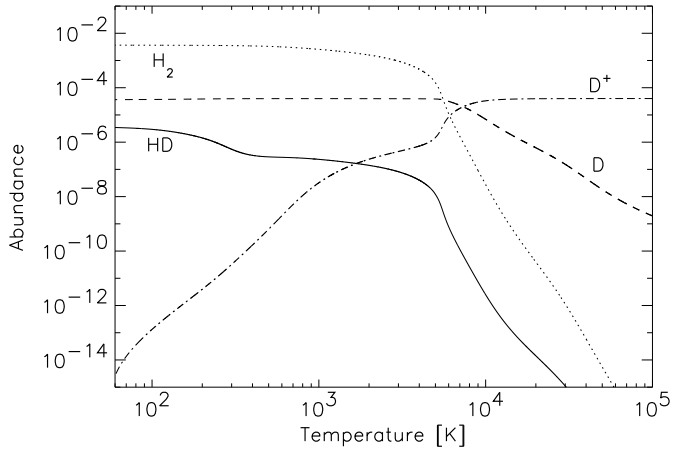

Figure 5.10 Molecular abundances in primordial gas cooling from high temperatures, relative to the total number density of H atoms. The calculation here simulated cooling in a 100 km s^{-1} shock at $z = 20$, characteristic of a supernova. Note the large abundance of HD at low temperatures. Figure credit: Johnson, J. L. & Bromm, V., *MNRAS*, **366**, 247 (2006).

Table 5.2 Reaction rates for Deuterium species as functions of temperature T in K [with $T_\xi \equiv (T/10^\xi \mathrm{K})$].

	Reaction	Rate Coefficient (cm^{3}s^{-1})
(1)	D$^+$ + e^- → D + $h\nu$	$8.40 \times 10^{-11} T^{-1/2} T_3^{-0.2} (1 + T_6^{0.7})^{-1}$
(2)	D + H$^+$ → D$^+$ + H	$3.70 \times 10^{-10} T^{0.28} \exp(-43/T)$
(3)	D$^+$ + H → D + H$^+$	$3.70 \times 10^{-10} T^{0.28}$
(4)	D$^+$ + H$_2$ → H$^+$ + HD	2.10×10^{-9}
(5)	HD + H$^+$ → H$_2$ + D$^+$	$1.00 \times 10^{-9} \exp(-464/T)$

excited state are

$$\frac{n_1}{n_0} = 3e^{-T_D/T}, \tag{5.41}$$

where the ratio of statistical weights is 3 and $T_D = h\nu_{10}/k_B \approx 130\,\mathrm{K}$ is the equivalent temperatures for photons emitted in transitions from the first excited state to the ground state. (We will take a two-level system for simplicity, assuming that the gas has already cooled to $T \sim T_D$ so that higher levels are rare.)

The radiative cooling rate of gas at constant density is

$$h\nu_{10}(n_0 B_{01} I_{\nu_{10}} - n_1 A_{10} - n_1 B_{10} I_{\nu_{10}}) = \frac{3}{2} n k_B \frac{dT}{dt}. \tag{5.42}$$

Here B_{01} and B_{10} are the Einstein coefficients for stimulated emission and absorption, respectively. n is the *total* number density of particles: this is related to the density of HD molecules $n_{\rm HD} = n_0 + n_1$ by $f_{\rm HD} = n_{\rm HD}/n$. Finally, $I_{\nu_{10}}$ is the CMB intensity at the frequency of the HD $J = 1 \rightarrow 0$ transition,

$$I_{\nu_{10}} \approx \frac{2h\nu_{10}^3}{c^2} e^{-T_D/T_{\rm CMB}} = \frac{A_{10}}{B_{10}} e^{-T_D/T_{\rm CMB}}, \tag{5.43}$$

where we have used the fact that $T_D \gg T_{\rm CMB}$. In that case the stimulated emission term can also be neglected, so equation (5.42) may be written

$$\frac{dT}{dt} \approx 2T_D A_{10} X_{\rm HD} (e^{-T_D/T_{\rm CMB}} - e^{-T_D/T}). \tag{5.44}$$

If we assume that $X_{\rm HD}$ remains constant, we can integrate this equation to find that the time to cool from $T \sim T_D$ to $T = T_{\rm CMB}$ is

$$t_{\rm HD,cool} \sim 1/(X_{\rm HD} A_{10}), \tag{5.45}$$

at $z \sim 10$–30. Equating this to the Hubble time, we can determine the critical HD abundance for cooling as

$$X_{\rm HD,crit} \sim 4 \times 10^{-9} \left(\frac{1+z}{30}\right)^{3/2}. \tag{5.46}$$

Figure 5.10 (and similar calculations for other scenarios) show that, when cooling from high temperatures, the gas forms far more HD than this critical value, indicating very efficient HD cooling. Since the Big Bang nucleosynthesis expectation is that the deuterium abundance is only $\sim 10^{-5}$ that of hydrogen, these calculations indicate that nearly all of the deuterium can enter molecular form. On the other hand, the abundance of HD in the "normal" Population III.1 scenario is well below this critical value – because D^+ is so rare in cold gas – so it is not an important coolant for that star formation channel.

5.3.3 The Population III.2 IMF

The previous section showed that the characteristic temperature of star-forming gas in this channel is much smaller than for Population III.1 stars, with $T \sim T_{\rm CMB}$. Such an effective cooling will lead to Bonnor-Ebert masses of ~ 10–$50\ M_\odot$, depending on the physics of cooling (see §5.2.4). This likely limits the masses of these

Population III.2 stars to be just a few tens of solar masses, considerably below the upper limits on Population III.1.

Numerical simulations show that a small protostar (with $m_\star < 0.5\ M_\odot$) forms in the Population III.2 case, just as in the case without deuterium, and subsequent stages proceed similarly to that case as well. However, in the colder gas, fragmentation into smaller mass protostars is much more likely, and the protostars are very unlikely to grow to the $\sim 100\ M_\odot$ scales necessary to make radiative feedback relevant. Thus it appears plausible that the Population III.2 IMF is skewed toward high-mass stars, but stars that still lie within the mass range observed in the nearby Universe.

This second-generation process may therefore produce a much different IMF than the first generation. However, we have seen that turbulence, chemical processes, and gravitational instability may cause even Population III.1 protostellar systems to fragment into clumps of comparable sizes. It remains to be seen how different these two formation channels really are.

5.4 PROPERTIES OF THE FIRST STARS

If fragmentation is inefficient, Population III stars appear to grow many times more massive than the Sun, probably ceasing accretion only when radiative feedback becomes important (§5.2.2). Primordial stars with $m_\star > 100 M_\odot$ have an effective surface temperature $T_{\rm eff}$ approaching $\sim 10^5$ K, with only a weak dependence on their mass. This temperature is ~ 17 times higher than the surface temperature of the Sun, ~ 5800 K. These massive stars are held against their self-gravity by radiation pressure, having the so-called *Eddington luminosity* (see Eq. 5.27 above, and the discussion in §7.3) which is strictly proportional to their mass $m_\star$,

$$L_E = 1.3 \times 10^{40} \left(\frac{m_\star}{100 M_\odot} \right)\ {\rm erg\ s^{-1}}, \tag{5.47}$$

and is 6–7 orders of magnitude more luminous than the Sun, $L_\odot = 4 \times 10^{33}\ {\rm erg\ s^{-1}}$. Because of these characteristics, the total luminosity and color of a cluster of such stars simply depends on its *total* mass and not on the mass distribution of stars within it.

The radii of these stars $R_\star$ can be estimated by equating their luminosity to the emergent blackbody flux $\sigma T_{\rm eff}^4$ times their surface area $4\pi R_\star^2$ (where $\sigma = 5.67 \times 10^{-5}\ {\rm erg\ cm^{-2}\ s^{-1}\ deg^{-4}}$ is the Stefan-Boltzmann constant). This gives

$$R_\star = \left(\frac{L_E}{4\pi\sigma T_{\rm eff}^4} \right)^{1/2} \approx 4.3 \times 10^{11}\ {\rm cm} \times \left(\frac{m_\star}{100 M_\odot} \right)^{1/2}, \tag{5.48}$$

which is only mildly larger than the radius of the Sun, $R_\odot = 7 \times 10^{10}$ cm.

The high surface temperature of the first stars makes them ideal factories of ionizing photons: liberating the electron from hydrogen requires an energy of 13.6 eV, while helium requires 24.4 eV for the first electron and 54.6 eV for the second. These are coincidentally near the characteristic energy of a photon emitted by these very massive Population III stars.

If indeed they were this massive, the first stars had lifetimes of a few million years, independent of their mass, because $L \propto m_\star$. During its lifetime, a very massive Population III star produced $\sim 10^5$ ionizing photons per proton incorporated in it; the precise efficiency depends on mass and the model parameters of the star, but only to within a factor of ~ 2 in the $m_\star = 10^2$–$10^3\ M_\odot$ range. This means that only a tiny fraction ($> 10^{-5}$) of all the hydrogen in the Universe needs to be assembled into Population III stars in order for there to be sufficient photons to ionize the rest of the cosmic gas, a fact which may be important during the reionization process (see chapter 8). For comparison, Population II stars with a standard Salpeter IMF (eq. **??**) produce on average $\sim 4,000$ ionizing photons per proton in them.

If fragmentation is permitted, the masses may be considerably smaller – > 10–$50\ M_\odot$, much larger than the characteristic mass today but still within the range of "normal" stars. In this case, the Population III stars will not be qualitatively different from their present-day analogs, although there are of course some differences in detail.

Evolutionary models of Population III stars are fairly well-specified, with the primary uncertainty at the high-mass end being the degree of mass loss during the stellar evolution. Figure 5.11 show some example calculations. The solid lines show main-sequence evolutionary tracks for zero metallicity stars without mass loss, while the short-dashed lines assume strong mass loss. Similar evolutionary tracks are shown for low-metallicity stars ($Z = 0.02\ Z_\odot$) with the dotted lines, and the zero-age main sequence for solar-metallicity stars is shown with the vertical solid line. Primordial stars tend to be hotter (or bluer) than their enriched counterparts (as well as slightly smaller). There are two reasons for this. First, the CNO cycle is inefficient (only able to use the small amount of carbon built up during the pre-main sequence phase). They thus have very hot cores. The lack of heavy elements also reduces the opacity of the outer layers. Together these factor imply hotter stellar surfaces.

These lower mass stars are therefore somewhat more efficient at producing ionizing photons than Population II (or I) stars, but the difference is one of quantity rather than quality, emitting roughly $\sim 50\%$ more ionizing photons per unit mass. The overall efficiency of producing ionizing photons will therefore depend extremely sensitively on the IMF: only if very massive stars are indeed able to form will Population III stars be orders of magnitude more efficient than later generations of stars. Figure 5.12 illustrates this very important point: it shows the observed spectrum of two Population III star clusters, one with purely very massive stars (solid line; in this case the spectrum is mostly independent of the mass distribution of the stars) and a standard Salpeter IMF (dotted line). For the same total stellar mass, the observable flux is larger by an order of magnitude for stars biased towards having masses $> 100 M_\odot$.

5.4.1 Emission Lines: Signatures of Primordial Stars

The hotter temperatures and increased ionizing efficiencies of massive Population III stars imply that galaxies in which massive stars are prevalent will have some

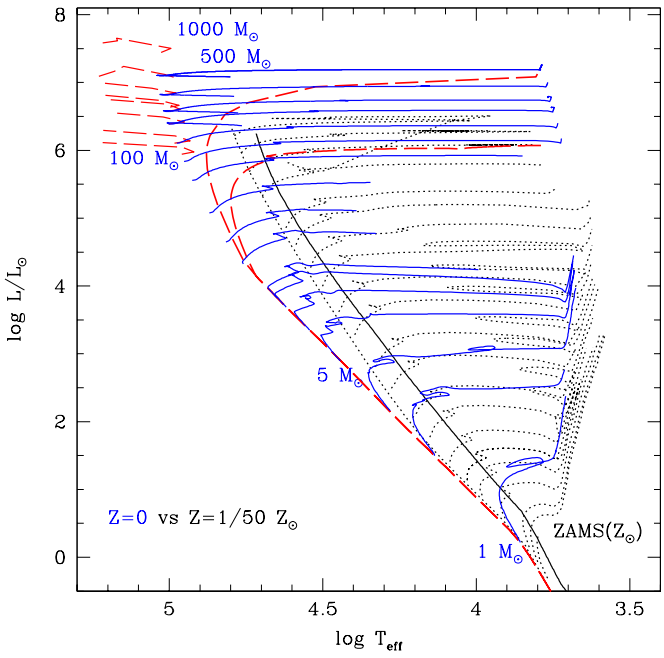

Figure 5.11 Main sequence evolutionary tracks for Population III stars (solid lines: without mass loss; short-dashed lines: with strong mass loss) and $Z = 0.02\ Z_{\odot}$ stars (dotted lines). Isochrones at 2 and 4 Myr for the $Z = 0$ stars are also shown with the long-dashed lines. The zero-age main sequence for solar-metallicity stars is shown by the vertical solid line; note that Popuation III stars are significantly hotter (bluer) than their higher-metallicity counterparts. Figure credit: Schaerer, D., *A & A*, **382**, 28 (2002).

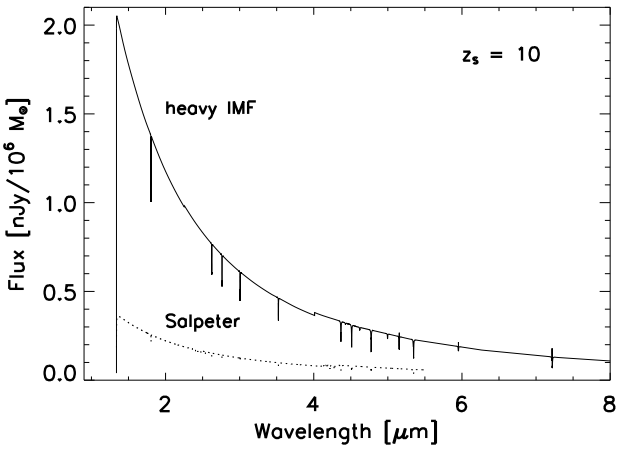

Figure 5.12 Comparison of the observed flux per unit frequency from a cluster of Population III stars at a redshift $z_s = 10$ for a Salpeter IMF (*dotted line*) and an IMF composed purely of very massive stars (*solid line*). The flux in units of nJy per $10^6 M_\odot$ of stars is plotted as a function of observed wavelength in μm. The cutoff below an observed wavelength of $1216\,\text{Å}\,(1+z_s) = 1.34\mu$m is due to hydrogen Lyman-α absorption in the IGM (the so-called Gunn-Peterson effect; see §10.2). Figure credit: Bromm, V. Kudritzki, R. P. & Loeb, A. *Astrophys. J.* **552**, 464 (2001); Salpeter curve from Tumlinson, J., & Shull, M. J. *Astrophys. J.* **528**, L65 (2000).

interesting observational signatures. As the high-energy photons escape into the interstellar media of their host galaxies, many of those photons will encounter neutral hydrogen or helium and be absorbed. The ionized gas will then recombine, emitting one or more line photons as the atom returns to the ground state. The relative numbers of these line photons depend on the incident spectra and so can be used as diagnostics of the stellar IMF.

Let us define $\dot{Q}_{i,\star}$ as the rate at which a star of mass m produces photons capable of ionizing a species i. Because line emission is the result of absorbing these photons, we have for a line m

$$L_m = f_m h\nu_m (1 - f_{\rm esc})\dot{Q}_{i,\star}, \tag{5.49}$$

where $f_{\rm esc}$ is the fraction of photons that escape the galaxy without absorption, the choice of species i depends on the transition m, $h\nu_m$ is the energy of a photon emitted in transition m, and f_m describes how likely a recombination of the appropriate species is to produce a photon in this line. Because these last two factors depend only on atomic physics, the ratios of different lines provide the ratios of ionizing photons and hence a measure of the spectral hardness of the local stellar population, albeit modulated by the factor $(1 - f_{\rm esc})$, which could in principle also depend on frequency.

In fact, if all of these ionizing photons are absorbed within the host galaxy, the hot, dense nebulae create substantial continuum emission as well, through free-free emission from the hot electrons, free-bound emission (by H I, He I, and He II) from the recombinations themselves, and the two-photon continuum of H I (generated when atoms recombine through the $2S$ level, which is metastable but eventually decays to the ground state by emitting two photons; see §11.2.2 for more on this process). This redistributes a large fraction of the energy contained in ionizing photons to lower energies and can substantially boost the brightness of the galaxies.

Figure 5.13 shows an example spectrum of a zero-age Population III star cluster, in which the IMF contains high-mass stars but is not exclusively made up of them. The solid curve shows the spectrum including the reprocessing from nebulae and recombination lines; the long-dashed curve shows the stellar continua themselves. Because such a large fraction of the energy is originally invested in ionizing photons, this reprocessing enhances the the rest-optical continuum by nearly an order of magnitude, and creates very strong lines. Here H I lines are shown with solid lines, He I with short-dashed lines, and He II with long-dashed lines.

The dotted curve shows the spectrum of a Population II cluster with $Z = 0.02\,Z_\odot$ and a Salpeter IMF ranging from 1–150 $M_\odot$ (normalized to the same total mass). The Population III case is somewhat brighter. More striking is the presence of the He II recombination lines at 1640, 3203, and 4868 Å, which appear because the highest mass stars are so hot and so produce a substantial amount of energy (up to $\sim 12\%$) above the He II ionization edge. In standard models, higher metallicity (and hence lower mass) stars produce almost no photons above this level, so these recombination lines are very interesting signatures of very massive Population III stars. (Though these lines *do* appear in Wolf-Rayet stars and in some star-forming galaxies at lower redshifts.)

However, because the highest mass stars live for only a few Myr, these He II

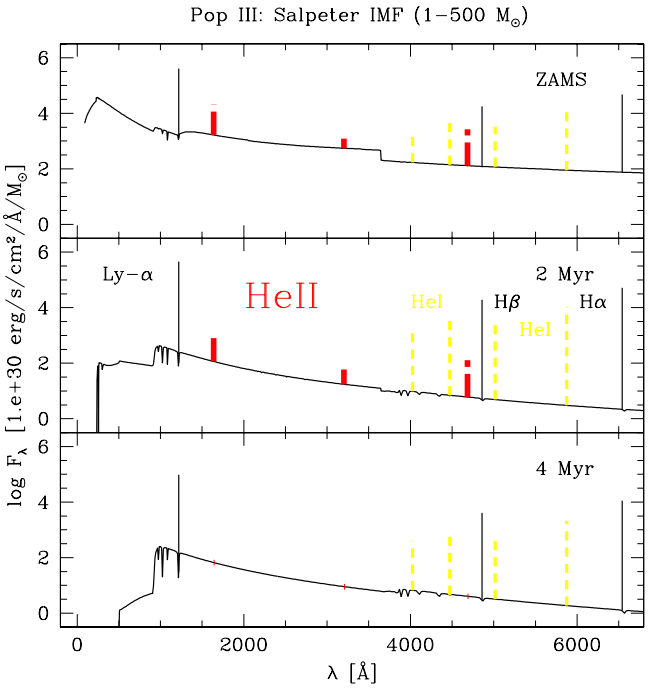

Figure 5.13 Spectral energy distribution of a cluster of Population III stars with a Salpeter IMF ranging from 1–500 $M_\odot$ (solid line), all of which have just entered the main sequence. Nebular reprocessing and recombination line emission are included assuming that $f_{\rm esc} = 0$. The pure stellar continuum (neglecting nebular emission) is shown by the long-dashed line. The contrasting case of a Population II cluster with $Z = 0.02\ Z_\odot$ and a Salpeter IMF ranging from 1–150 $M_\odot$ is shown by the dotted line. The vertical dashed lines indicate the ionization potentials of H I, He I, and He II (from right to left). Figure credit: Schaerer, D., *A & A*, **382**, 28 (2002).

recombination lines do not persist for long after an initial burst of star formation. They are therefore not *necessary* signatures of zero-metallicity stars, even if they are convenient markers.

5.5 THE END STATES OF POPULATION III STARS

The end state in the evolution of massive Population III stars depends on their mass and rotation rate. Ignoring rotation, one finds several possible fates, depending on the initial stellar mass, though modeling supernovae is extremely difficult so the dividing lines between the different scenarios remain uncertain. Rotation generally increases the mass thresholds identified below, but it has proven difficult to quantify by how much. We describe each of these fates briefly below.

- For masses below ~ 8–$10\ M_\odot$, stars end their lives as white dwarfs, just as present day low-mass stars do. These stars can produce light elements during their asymptotic giant branch phases, but that occurs over much longer timescales than the < 1 Gyr Hubble time at $z > 6$, so it is generally not considered important in understanding the enrichment histories of early galaxies.

- For masses ~ 10–$25\ M_\odot$, stars undergo Type II supernovae, leaving a neutron star behind. Especially at low metallicities, where the opacities are smaller, the hydrogen envelopes remain intact: these are the "normal" supernovae that are thought responsible for enrichment of very heavy elements in the nearby Universe.

- For masses ~ 25–$40\ M_\odot$, stars undergo relatively weak Type II supernovae because much of the ^{56}Ni falls back onto the black hole remnant. As a result, these supernovae are likely quite faint and leave little iron behind.

- For masses ~ 40–$100\ M_\odot$, the stars collapse directly to a black hole *without* producing a supernova (and hence without enriching their surroundings).

- For masses between 100–$140\ M_\odot$, the enormous core following helium burning heats up rapidly, leads to the production of electron-positron pairs as a result of collisions between atomic nuclei and energetic gamma-rays, which in turn reduces thermal pressure inside the star's core. This instability creates violent mass-ejecting pulsations, which can contain as much energy as a supernova (though will be much fainter due to the lack of radioactive elements). The entire hydrogen envelope of the star is likely ejected, relieving the instability and allowing the remainder of stellar evolution to proceed as for a lower-mass star, with the iron core eventually collapsing directly to a black hole. These kinds of "explosions" would not enrich their galaxies because only the light envelopes are ejected.

- For the mass range 140–$260\ M_\odot$, stars are likely to explode as *pair-instability supernovae*. A pair-instability supernova is triggered by the same instability described above, when part of the core's thermal energy is invested in the rest

mass of electron-positron pairs. The pressure drop leads to a partial collapse and then greatly accelerated burning in a runaway thermonuclear explosion which blows the star up without leaving a remnant behind. The kinetic energy released in the explosion could reach $\sim 10^{53}$ ergs, exceeding the kinetic energy output of typical supernovae by two orders of magnitude. Although the characteristics of these powerful explosions were predicted theoretically several decades ago, there has been no conclusive evidence for their existence so far. Because of their exceptional energy outputs, pair-instability supernovae would be prime targets for future surveys of the first stars with the next generation of telescopes (§9.1.2). Because of their unusual explosion mechanisms, pair instability supernovae have distinct nucleosynthetic signatures. They produce a near solar distribution of elements from oxygen to nickel *except* with a large deficit of nuclei with odd charges, because weak interactions are unimportant throughout most of this mass range. They are also unable to make very heavy elements and eject no elements heavier than zinc.

- Above $260\ M_{\odot}$, the helium cores instead collapse directly to black holes; nuclear burning of heavier elements is simply unable to halt the implosion triggered by exhaustion of more efficient fuel, and the entire star is swallowed up in the black hole, though possibly with a transient accretion disk and accompanying electromagnetic signature. Above this mass, Population III stars therefore do *not* enrich their surroundings.

5.6 GAMMA-RAY BURSTS: THE BRIGHTEST EXPLOSIONS

Gamma-ray bursts (GRBs) were discovered in the late 1960s by the American Vela satellites, built to search for flashes of high energy photons ("gamma rays") from Soviet nuclear weapon tests in space. The United States suspected that the Soviets might attempt to conduct secret nuclear tests after signing the Nuclear Test Ban Treaty in 1963. On July 2, 1967, the Vela 4 and Vela 3 satellites detected a flash of gamma radiation unlike any known nuclear weapons signature. Uncertain of its meaning but not considering the matter particularly urgent, the team at the Los Alamos Laboratory, led by Ray Klebesadel, filed the data away for future investigation. As additional Vela satellites were launched with better instruments, the Los Alamos team continued to find unexplained GRBs in their data. By analyzing the different arrival times of the bursts as detected by different satellites, the team was able to estimate the sky positions of 16 bursts and definitively rule out either a terrestrial or solar origin. The discovery was declassified and published in 1973 (*Astrophys. J.* **182**, L85) under the title "Observations of Gamma-Ray Bursts of Cosmic Origin."

The distance scale and nature of GRBs remained mysterious for more than two decades. Initially, astronomers favored a local origin for the bursts, associating them with sources within the Milky Way. In 1991, the Compton Gamma Ray Observatory satellite was launched, and its "Burst and Transient Source Explorer"

instrument started to discover a GRB every day or two, increasing the total number of known GRBs up to a few thousand. The larger statistical sample of GRBs made it evident that their distribution on the sky is isotropic. Such a distribution would be most natural if the bursts originate at cosmological distances since the Universe is the only system which is truly isotropic around us. Nevertheless, the local origin remained more popular within the GRB community for six years, until February 1997, when the Italian-Dutch satellite BeppoSAX detected a gamma-ray burst (GRB 970228) and localized it to within minutes of arc using its X-ray camera. With this prompt localization, ground-based telescopes were able to identify a fading counterpart in the optical band. Once the GRB afterglow faded, deep imaging revealed a faint, distant host galaxy at the location of the optical afterglow of the GRB. The association of a host galaxy at a cosmological distance for this burst and many subsequent ones revised the popular opinion in favor of associating GRBs with cosmological distances. This shift in popular view provides testimony to how a psychological bias in the scientific community can be overturned by hard scientific evidence.

A GRB afterglow is initially brightest at short photon wavelengths and then fades away at longer wavelengths, starting in the X-ray band (over timescales of minutes to hours), shifting to the UV and optical band (over days), and ending in the infrared and radio (over weeks and months).[ii] Among the first detected afterglows, observers noticed that as the afterglow lightcurve faded, long-duration GRBs showed evidence for a supernova flare, indicating that they are also associated with core-collapse supernova events. The associated supernovae were classified as related to massive stars which have lost their hydrogen envelope in a wind. In addition, long-duration GRBs were found to be associated with star-forming regions where massive stars are born and explode only a million years afterwards. These clues indicated that long-duration GRBs are most likely associated with massive stars. The most popular model for long-duration GRBs became known as the "collapsar" model. According to this model, the progenitor of the GRB is a massive star whose core eventually consumes its nuclear fuel, loses pressure support, and collapses. If the core of the star is too massive to make a neutron star, it collapses to a black hole. As material is spiraling into the black hole, two jets are produced at a speed close to that of light. So far, there is nothing spectacular about this setting, since we see scaled-up versions of such jets being formed around massive black holes in the centers of galaxies, as shown in Figure 7.5. However, when jets are generated in the core of a star, they make their way out by drilling a hole in the surrounding dense envelope. As soon as the head of a jet exits, the highly collimated stream of radiation emanating from it would appear as a gamma-ray flash to an observer who happens to line up with the jet axis. The subsequent afterglow results from the interaction between the jet and the ambient gas in the vicinity of the progenitor star. As the jet slows down by pushing against the ambient medium, the non-thermal radiation from accelerated relativistic electrons in the shock wave in front of it gets shifted to longer wavelengths and fainter luminosities. Also, as the jet makes its

[ii] For an extreme example of a GRB afterglow from a redshift $z = 0.94$ that was bright enough to be seen with the naked eye, see Bloom, J., et al. *Astrophys. J.* **691**, 723 (2009).

way out of the star, its piston effect deposits energy in the stellar envelope and explodes the star, supplementing the GRB with a supernova-like explosion. Because of their immense luminosities, GRBs can be observed out to the edge of the Universe. These bright signals may be thought of as the cosmic fireworks signaling the birth of black holes at the end of the life of their parent massive stars. If the first stars produced GRBs (as their descendants do in the more recent Universe), then they would be detectable out to their highest redshifts. Their powerful beacons of light can be used to illuminate the dark ages and probe the cosmic gas around the time when it condensed to make the first galaxies. As this book was written, a gamma-ray burst was discovered by the Swift Satellite[iii] at a redshift 9.4, representing the most distant source known, originating at the time when the Universe was only ~ 0.5 billion years old.

It is unknown whether Population-III stars produce long-duration GRBs. For that to happen, the angular momentum of the collapsing core mass M_c needs to be larger than $\sim 10GM_c^2/c$ so that a stable disk would form outside the resulting black hole and collimate the jets. The rotation of the pre-GRB progenitor can be affected by mass exchange with a binary companion or mass loss through a wind. If the final mass of the black hole from a Population III progenitor is larger than usual, then the duration and total energy output of the associated GRB is expected to increase ($\propto m_\star$) relative to low redshift GRBs. For additional observational details about GRBs, see §???.

[iii] http://swift.gsfc.nasa.gov/

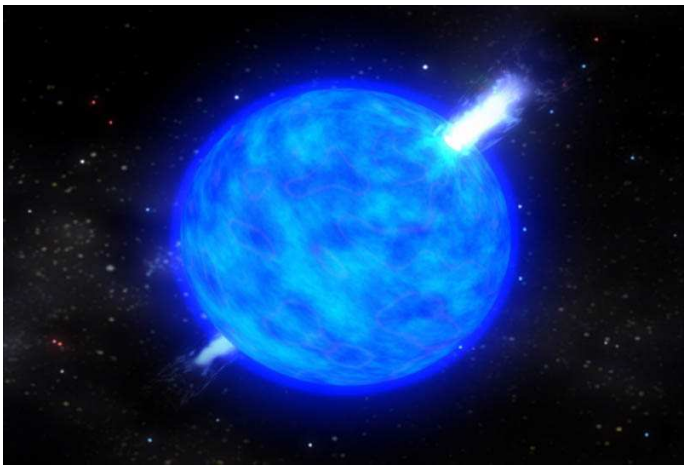

Figure 5.14 Illustration of a long-duration gamma-ray burst in the popular "collapsar" model. The collapse of the core of a massive star (which lost its hydrogen envelope) to a black hole generates two opposite jets moving out at a speed close to the speed of light. The jets drill a hole in the star and shine brightly towards an observer who happens to be located within the collimation cones of the jets. The jets emanating from a single massive star are so bright that they can be seen across the Universe out to the epoch when the first stars formed. Figure credit: NASA E/PO.

Chapter Six

Stellar Feedback and Galaxy Formation

6.1 THE ULTRAVIOLET BACKGROUND AND H_2 PHOTODISSOCIATION

Chapter 5 described star formation in gas with a primordial composition. We found that this process crucially depends on molecular hydrogen to cool the cloud to densities high enough for stars to form. In this section we will consider how radiation from those very same stars can destroy that coolant and so make subsequent star formation even harder.

6.1.1 Lyman-Werner Photons and the Solomon Process

Molecular hydrogen (H_2) is fragile and can easily be photodissociated by photons with energies of 11.5–13.6eV, to which the intergalactic medium (IGM) is transparent even before it is ionized. The photodissociation occurs through a two-step process, first suggested by Phil Solomon in 1965 and later analyzed quantitatively by Stecher & Williams (1967). In practice, this process is the only way to photodissociate H_2 in interstellar (or intergalactic) space, because the photodissociation continuum of H_2 begins at 14.7 eV, while the photoionization continuum begins at 15.4 eV.[i] Both of these lie above the photoionization threshold of H I, so such photons would be absorbed by H I long before they encountered H_2.

The quantum mechanical configuration of the electronic ground state of H_2 is denoted $X^1\Sigma_g^+$. Uppercase Greek letters denote the total electronic angular momentum of the system, projected onto the internuclear axis, with Σ, Π, and Δ having values of 0, 1, and 2 in units of $\hbar$. The left superscript (1 for the ground state) is $2S+1$, where $S=0$ or 1 is the total spin angular momentum. The right subscript (g or u) and superscript ($+$ or $-$, and only for Σ states) describe the symmetries of the configuration; this one asymptotes to two atoms with their electrons in the $1s$ state at large separations. The leftmost Roman letter describes the electronic states, with X being the lowest level, with the relevant upper states for our purposes labeled B and C (capitalized letters refer to singlets). Each of these electronic states is further split into a large number of sublevels by the quantized rotational and vibrational levels of the two nuclei, usually denoted by N and v. For example, the ground state has 14 vibrational levels, each nominally with an infinite number of rotational levels.

The next two singlet states are $B^1\Sigma_g^+$ and $C^1\Pi_u$, which asymptote to two atoms with their electrons in the $1s$ and $2s$ or $2p$ states, respectively. These can decay

[i] These are far above the dissociation energy of H_2 (4.48 eV) because the direct transition is forbidden.

to the ground state via permitted electric dipole transitions, the analog of the H I Lyman-α transition. However, for these molecules there are a large number of sub-transitions owing to the rotational and vibrational splittings. Thus, H_2 has two *bands* representing these transitions. The first band between the ground state and $B^1\Sigma_g^+$ is known as the *Lyman* band and consists of many densely packed lines beginning at 1108 Å(11.26 eV). The second band between the ground state and $C^1\Pi_u$ is known as the *Werner* band and begins at 1040 Å(12.3 eV).

Now consider the following sequence:

$$\mathrm{H}_2(X^1\Sigma_g^+, v=0) + \gamma \rightarrow \mathrm{H}_2(B^1\Sigma_g^+, v=v') \rightarrow \mathrm{H}_2(X, v=v'') + \gamma \quad (6.1)$$

Here v labels the vibrational energy level. Crucially, in electronic transitions there are no sharp selection rules for the vibrational continuum. Thus, the excited state's vibrational quantum number v' is not restricted to be small, and nor is the final value v''. It is therefore possible for the final state to lie in the vibrational continuum of the molecule ($v'' > 14$): in other words, to dissociate the molecule. A similar process also occurs for excitations and decays through the Werner band. The rate at which this occurs depends on the cross-section for absorbing Lyman-Werner photons (for which the oscillator strengths are typically $\sim 1\%$) and the probability of decay into this continuum (typically $\sim 15\%$). Figure 6.1 shows the energies of some of these transitions, where the initial configuration has $v = 0$ and $J = 0, 1$; the height of each line is $0.01 \times f_{\rm osc}$. The *average* cross-section for this process between 11.26 eV and 13.6 eV (averaged over 76 allowed lines) is $\sigma_{\rm LW} = 3.71 \times 10^{-18}\ {\rm cm}^2$.

6.1.2 The Suppression of H_2 Cooling

Once Lyman-Werner photons appear, we must include this photodissociation process in the chemistry of the primordial clouds. The rate coefficient for photodissociation is

$$k_{\rm diss} = 1.38 \times 10^9 J_{\rm LW}\ {\rm s}^{-1}, \quad (6.2)$$

where $J_{\rm LW}$ is the specific intensity (in units ergs s^{-1} cm^{-2} Hz^{-1} sr^{-1}) in the Lyman-Werner band (specifically, here we have taken $h\nu = 12.87$ eV for concreteness, in the middle of the relevant energy range). It is convenient to normalize $J_{\rm LW} = 10^{-21} \times J_{\rm LW,21}$. The timescale for dissociation is therefore

$$t_{\rm diss} = k_{\rm diss}^{-1} \approx 3 \times 10^4 J_{\rm LW,21}^{-1}\ {\rm yr}, \quad (6.3)$$

which is very short compared to the relevant cosmological timescales. Thus, if the Lyman-Werner background approaches this fiducial value, we would expect it to destroy all of the molecular hydrogen.

In that case, if the radiation background (and local gas properties) remain constant on longer timescales, the H_2 fraction will approach an equilibrium where the formation rate (approximately $\tilde{k}$ in equation 5.6) balances the dissociation rate,

$$f_{\rm H_2,eq} = \frac{\tilde{k}}{k_{\rm diss}} x_{\rm HII} n \sim 4 \times 10^{-8} J_{\rm LW,21}^{-1} \left(\frac{x_{\rm HII}}{2\times10^{-4}}\right) \left(\frac{1+z}{20}\right)^3 \left(\frac{\Delta}{200}\right), \quad (6.4)$$

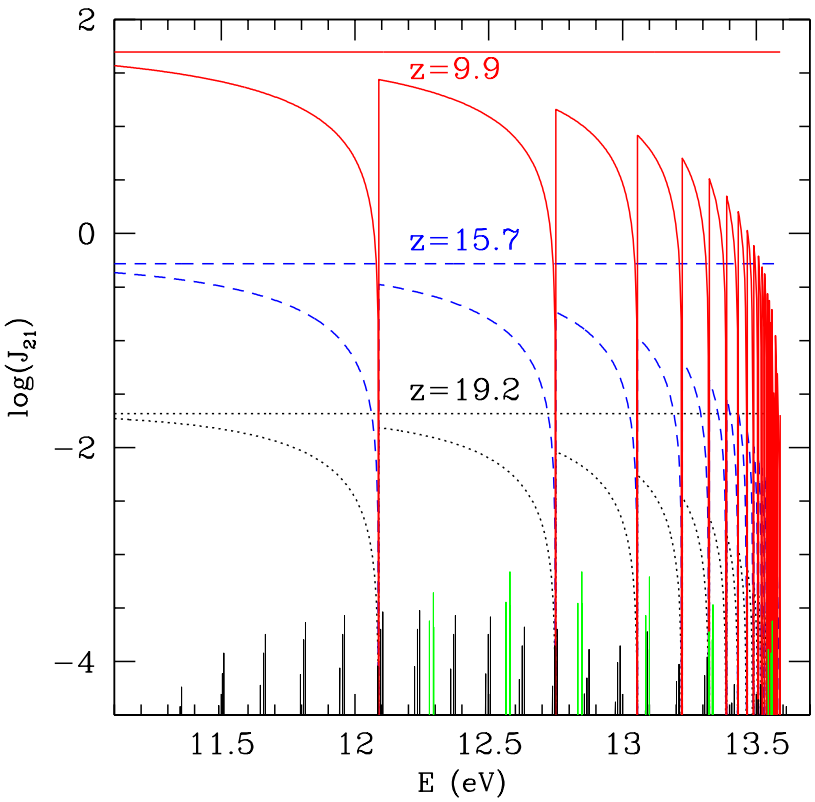

Figure 6.1 The "sawtooth" modulation of a uniform, spectrally flat radiation background in the Lyman-Werner frequency band when the IGM is still predominantly neutral. The three curves are for $z = 19.2$, 15.7, and 9.2, from top to bottom; the horizontal lines show the unattenuated spectrum, while the curves with features show the effect of Lyman-series absorption. The vertical lines at the bottom of the figure show some of the Lyman-Werner transitions, with the height equal to 1% of the oscillator strength. Figure credit: Ahn, K. et al., *Astrophys. J.* **695**, 1430 (2009).

where we have taken $T \sim 1000$ K and a typical electron fraction before any cooling begins. This is far below the critical value required for H_2 cooling to be efficient (Eq. 5.12), so *a substantial Lyman-Werner background suppresses molecular hydrogen cooling*.

The primary question is then whether a background of this amplitude can reasonably penetrate the clouds in which primordial stars may form. In the next section we will examine whether a sufficiently strong background can be produced by the integrated stellar population, but before that we note that any such metagalactic radiation field must penetrate to the regions in which H_2 actually forms – that is, the centers of virialized halos. Once H_2 cooling becomes important, these halos have large masses of the gas, and the outer layers of each halo can then *self-shield* the inner layers where cooling is actually occurring. If these outer layers are dense enough to maintain an equilibrium H_2 population that is optically thick in the Lyman-Werner bands, this self-shielding is significant. A convenient numerical approximation, due to Draine & Bertoldi (1996), for the effects of self-shielding in a static medium is to take $k_{\rm diss} \rightarrow f_{\rm sh} k_{\rm diss}$, with

$$f_{\rm sh} = \min\left[1, \left(\frac{\rm N_{H_2}}{10^{14}\,{\rm cm}^{-2}}\right)^{-0.75}\right], \tag{6.5}$$

where $N_{\rm H_2}$ is the column density of molecular hydrogen. The dependence at high column densities is steeper than expected from a naive curve-of-growth analysis because of overlap within the various Lyman-Werner lines. (This estimate is not accurate at very high column densities, but those are rarely important in this context.)

Note, however, that self-shielding is more complex if the medium has velocity gradients, because then the lines are shifted by different amounts relative to their rest wavelengths in different parts of the cloud. This can considerably reduce the effectiveness of self-shielding and is a critical question in evaluating the importance of a Lyman-Werner background

6.1.3 Photodissociation Feedback Inside Star-Forming Halos

It is conceptually convenient to divide feedback from Lyman-Werner photodissociation into two simple cases: one in which starlight from a given star inside a collapsed halo acts upon gas inside the same halo, and second in which a metagalactic radiation background affects halos from the outside. We will consider the first case here. We certainly expect that, within some zone around an individual star, the Lyman-Werner background will dissociate enough H_2 to render further cooling inefficient, choking off later star formation. The question we wish to address is how large this zone is relative to the extent of the halo.

We suppose that a star sits at the center of such a halo. To gauge the cumulative amount of photodissociation, we must compare the timescale for a star to photodissociate the halo's H_2 to the main-sequence lifetime of the star. Very massive Population III stars produce $N_{\rm LW} \approx 3400$ photons in the 11.2–13.6 eV range per baryon inside them; smaller stars produce them at about double that rate. If we assume that a fraction $f_{\rm LW,abs} \sim 0.01$ of these photons are absorbed by the

Lyman-Werner bands (a reasonable approximation for the relevant column densities and expected line widths), and that about $f_{\rm LW,diss} \sim 0.15$ of these absorptions lead to dissociations, the total number of dissociations from a star (or set of stars) with mass $M_\star$ is $\sim f_{\rm LW,abs} f_{\rm LW,diss} N_{\rm LW} M_\star / m_p$. Comparing this to the total number of H_2 molecules in a halo, $\sim f_{\rm H_2} M_g / m_p$ (where M_g is the total gas mass), we find that the fraction of molecules expected to be photodissociated is

$$f_{\rm destroy} \sim 10^4 \left(\frac{f_{\rm LW,abs}}{0.01}\right) \left(\frac{f_{\rm LW,diss}}{0.15}\right) \left(\frac{N_{\rm LW}}{3400}\right) \left(\frac{f_{\rm H_2,s}}{3.5 \times 10^{-4}}\right)^{-1} \left(\frac{M_\star}{M_g}\right). \quad (6.6)$$

Thus, provided that the star formation efficiency is not extremely small, the first generation of stars can easily photodissociate all of their halo's diffuse H_2, shutting down further cooling at least temporarily.

However, gas clumps already in the process of collapse may already be dense enough to maintain their H_2 populations in the presence of this radiation background. The relevant question for clumps is whether the radiation field can dissociate the H_2 both before collapse completes (over $\sim t_{\rm dyn}$) and faster than the clump can form H_2 to replace it (Eq. 6.4). Analytic estimates show that clumps that have already passed the "loitering stage" (with $n > 10^6$ cm^{-3}) are suppressed only very close to the source star. Thus, the total rate of star formation within halos may depend on the degree to which clumps are synchronized across the entire halo: those collapsing at nearly the same time will be unaffected by the Lyman-Werner background, but the collapse of those that are delayed may be halted completely.

6.1.4 The Metagalactic Lyman-Werner Background

Because the intergalactic medium is mostly optically thin to photons in the Lyman-Werner bands (and the small amount of intergalactic H_2 is quickly dissociated as the first sources appear), a metagalactic radiation field will quickly build up in this energy range. If the background is intense enough, the rate at which H_2 is destroyed inside collapsed objects will exceed the rate at which such molecules form, preventing cooling in newly forming halos – and causing a strong *negative* feedback effect on star formation.

The magnitude of this feedback will depend upon how these Lyman-Werner photons propagate through the IGM. In fact, the IGM is not perfectly optically thin to these photons as line absorption by the H I Lyman series lines processes the background below the Lyman limit, causing the sawtooth shape shown in Figure 6.1. For any photon energy above Lyman-α at a particular redshift, there is a limited redshift interval beyond which no contribution from sources is possible because the corresponding photons are absorbed through one of the (extremely optically thick) Lyman-series resonances along the way.[ii] Consider, for example, an energy of 11 eV at an observed redshift $z = 10$. Photons received at this energy would have to be emitted at the 12.1 eV Lyman-β line from $z = 11.1$. Thus, sources in the redshift interval 10–11.1 could be seen at 11 eV, but radiation emitted by sources

[ii] The Lyman-α optical depth given in equation (4.13), and higher Lyman-series transitions are related to this fall only by the ratios of the oscillator strength times frequency.

at $z > 11.1$ eV would have passed through the 12.1 eV energy at some intermediate redshift, and would have been absorbed. An observer viewing the universe at any photon energy above Lyman-α would see sources only out to some horizon, and the size of that horizon would depend on the photon energy. The number of contributing sources, and hence the total background flux at each photon energy, would depend on how far this energy is above the nearest Lyman resonance. Most of the photons absorbed along the way would be re-emitted either at Lyman-α or in the $2p \rightarrow 1s$ two-photon continuum and then redshifted to lower energies. The result is a sawtooth spectrum for the UV background before reionization, with an enhancement below the Lyman-α energy.

Quantitatively, the specific intensity at a frequency ν and redshift z is

$$J_\nu(z) = \int dz' c \frac{dt}{dz'} j_{\nu'}(z') e^{-\tau(z)}, \tag{6.7}$$

where $j_{\nu'}(z')$ is the emissivity from sources at a redshift z' and a frequency $\nu' = \nu(1+z')/(1+z)$ and the factor $\tau(z)$ is the accumulated optical depth as the photon travels through the IGM. This is negligible so long as the photon stays between the Lyman-series lines, but it becomes very large whenever the photon crosses such a line. An excellent approach is therefore to use a "screening approximation" in which the integral is truncated at a maximum redshift determined by the nearest Lyman line i (of frequency $\nu_i > \nu$) via

$$\frac{1+z_{\rm max}}{1+z} = \frac{\nu_i}{\nu}, \tag{6.8}$$

while the optical depth factor can otherwise be ignored.

Figure 6.1 shows this modulation in detail for a set of uniform emissivity sources with flat spectra at three different final redshifts (the normalizations are arbitrary; the horizontal lines show the spectra before attenuation by the Lyman-series). As the frequency increases and the spacing between the Lyman-series lines decreases, the absorbing screens get closer together and the total background decreases. Thus, the uppermost Lyman-Werner transitions are affected by a weaker background.

Unfortunately, the direct detection of the redshifted sawtooth spectrum as a remnant of the reionization epoch is not feasible due to the much higher flux contributed by foreground sources at later cosmic times. However, a similar process does occur before He II is completely reionized at $z < 3$, with the Lyman-series transitions of He II creating a similar sawtooth spectrum in the far-ultraviolet. This may be indirectly detectable through its effects on metal-line absorbers, some of whose ionization potentials lay inside the sawtooth region of the spectrum.

Estimating the spectrum in more detail, and as a function of redshift, requires a model for the emissivity $j_\nu(z)$. Clearly that will depend on the galaxy formation processes that we will examine over the next several chapters, but for a very simple estimate we can assume that the star formation efficiency $f_\star$ within halos is zero below a minimum halo mass $M_{\rm min}$ and constant above that mass. Then we can write

$$j_\nu(z) = \frac{1}{4\pi} f_\star \frac{df_{\rm coll}}{dt} \frac{\bar{\rho}_b}{m_p} \epsilon_{\rm LW}(\nu), \tag{6.9}$$

where the first factor converts from total emissivity to emissivity per solid angle and the last factor is the energy produced by the stars per frequency per baryon in the Lyman-Werner region. Approximating the latter by $\epsilon_{\rm LW} \approx h\nu_{\rm LW} N_{\rm LW}/\Delta\nu_{\rm LW}$, equation (6.7) gives

$$J_{\nu,21} \sim 2.4 \left(\frac{N_{\rm LW}}{3400}\right) \left(\frac{f_\star}{0.1}\right) \left(\frac{\Delta f_{\rm coll}}{0.01}\right) \left(\frac{1+z}{10}\right)^3, \qquad (6.10)$$

where $\Delta f_{\rm coll}$ is the fraction of gas that collapses onto star-forming halos over the redshift range $(z, z_{\rm max})$. Radiation backgrounds of this magnitude are easily large enough to strongly suppress H_2 cooling in just-virialized gas (see Eq. 6.4).

Figure 6.2 shows a more careful calculation of the background spectrum amplitude, though still in the context of a model with the star formation rate proportional to $df_{\rm coll}/dt$ and $f_\star = 0.1$. Here we show the *average* amplitude over the entire Lyman-Werner frequency interval – the sawtooth absorption typically reduces this from the emitted amplitude by about an order of magnitude. We show several different mass thresholds, increasing from the filter mass (top curve) to masses near the atomic cooling threshold (bottom curve). The amplitude increases rapidly with decreasing redshift because these halos are initially on the exponential tail of the mass function; the turnover at lower redshifts is where the corresponding halos are well below the cutoff in the mass function so that the growth slows down. Equation (6.10) appears to provide a reasonable estimate of $J_{\rm LW,21}$.

The choice of $f_\star$ is highly uncertain in these models, so Figure 6.2 is only a very rough guide to expectations. If the first cluster of Population III.1 stars shuts down further star formation in a halo, then one might expect only a few hundred solar masses of stars to form. In that case, $f_\star = M_\star/M_g \sim 0.003(M_\star/500\,M_\odot)(M_h/10^6\,M_\odot)^{-1}$. Fortunately, these curves are all strictly proportional to that parameter, so that their amplitude can easily be rescaled.

This mean background is relatively easy to compute, but in reality the clustered halos that source the background induce inhomogeneities in it. Fortunately, at least in the standard structure formation model, these inhomogeneities are mild. Consider the lower edge of the Lyman-Werner band, with 11.2 eV. Photons redshift into this band out to the Lyman-β transition at 12.1 eV, which corresponds to a redshift of $\Delta z \sim 0.1(1+z)$, or about 100 comoving Mpc. Each point therefore samples a huge volume of sources around it (even though the more closely-spaced higher Lyman-series transitions weight the effective volume to more nearby sources), which averages out fluctuations. The Lyman-Werner background will therefore be nearly uniform except very close to individual sources or unless the halo population itself has fluctuations on ~ 100 Mpc scales, which may indeed be possible due to a strong source bias and baryon acoustic oscillations (see §3.3). In such a case, the background may vary strongly, leading to substantial variations in the halos able to cool and form stars efficiently.

6.1.5 External Feedback on H_2 Inside Virialized Halos

With a model for the Lyman-Werner background in hand, it is now straightforward to gauge the metagalactic background's effects on H_2 cooling inside collapsing

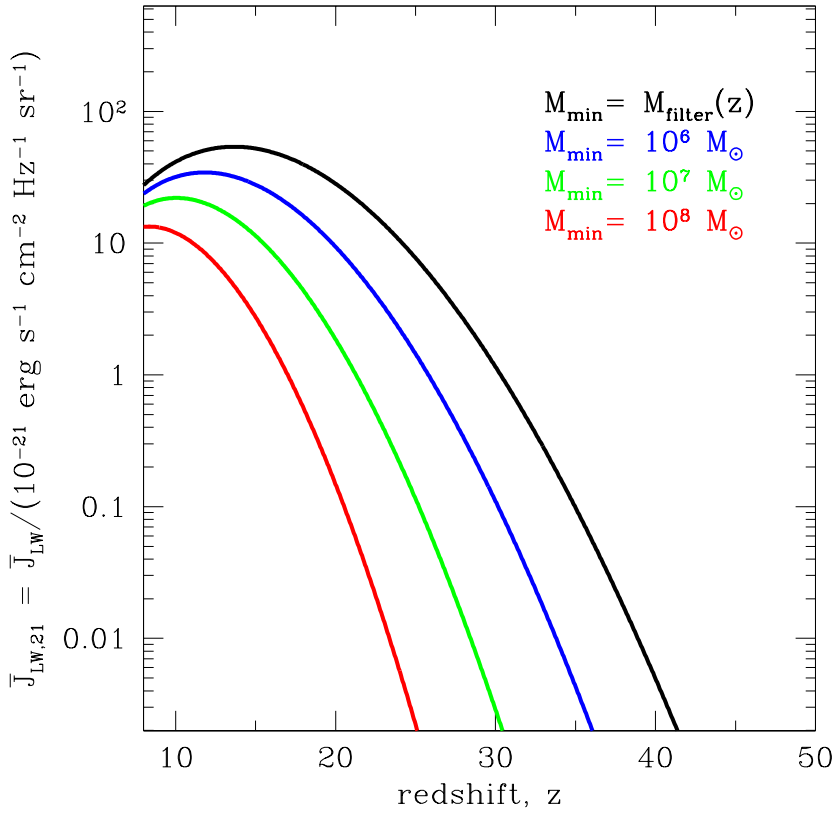

Figure 6.2 Evolution of the specific intensity of the metagalactic radiation field in the Lyman-Werner band at high redshifts. The solid lines show the amplitude of the radiation field over time, taking several different mass thresholds for star-forming halos: $M_{\rm min} = M_{\rm fil}$, 10^6, 10^7, and 10^8 $M_\odot$, from top to bottom. The curves assume $f_\star = 0.1$. Figure credit: L. Holzbauer (UCLA).

dark matter halos. As a simple estimate of the column density of a virialized halo, we assume a uniform density sphere at the typical virial overdensity and with a radius $r_{\rm vir}$. Then a halo of mass M has

$$N_{\rm H_2} \sim 10^{17} \left(\frac{f_{\rm H_2}}{3.5\times10^4}\right) \left(\frac{M}{10^6\ M_\odot}\right)^{1/3} \left(\frac{1+z}{20}\right)^2\ {\rm cm}^{-2}, \qquad (6.11)$$

where we have inserted the saturation value for the H_2 fraction from equation (5.10) as a fiducial estimate. In fact, simulations show that the effective column density is typically a few times smaller than this since much of the gas in the outskirts of the halo remains optically thin, but confirm that it provides a reasonable estimate for a stationary halo in which velocity gradients are insignificant.

This column density is well above the self-shielding threshold in equation (6.5), implying that much of the halo will be shielded from the metagalactic background. Therefore, we write the effective background as $f_{\rm sh} J_{\rm LW,21}(z)$. We can then insert this radiation field into equation (6.4) to determine the H_2 fraction in the presence of feedback. Finally, comparison of this fraction to the critical value required for cooling, $f_{\rm H_2,c}$ in equation (5.12) determines whether the halo is able to continue cooling and form stars.

Figure 6.3 provides a schematic illustration of these effects, based on fits to numerical simulations (c.f. Fig. **??**). The solid line shows $f_{\rm H_2,c}$, the critical fraction required for efficient cooling. The dashed line (marked *a*) shows $f_{\rm H_2}$ in the absence of radiative feedback; this lies very near the saturation level of equation (5.10). The thick dotted line (marked *b*) shows $f_{\rm H_2}$ if self-shielding is neglected and $J_{\rm LW,21} = 0.01$. This markedly reduces $f_{\rm H_2}$ and quantitatively matches the estimates described in this section. However, the dot-dashed line (marked *c*) shows the same, but with self-shielding approximately included. Halos near the critical cooling threshold are already very optically thick, so in practice the radiation background has substantially less of an effect than naively expected.

Nevertheless, the growing Lyman-Werner background will most likely "self-regulate" the earliest stages of star formation. Within each star-forming halo, the first few stars create a strong Lyman-Werner background and prevent any protostars not already far along in their collapse from proceeding. The same stars create a metagalactic background that reduces the efficiency of cooling in other, newly forming gas clouds, raising the mass threshold for star formation. But as the abundance and mass scale of dark matter halos increases, the gas clouds have better self-shielding and raise the background, which in turn raises the mass threshold, and so on. Eventually the Lyman-Werner background will become so intense that star formation is only possible through atomic cooling in halos with $T_{\rm vir} \sim 10^4$ K, for which photodissociation is unimportant. However, recall that these halos ionize their gas and so likely form stars through the Population III.2 (deuterium-mediated) channel described in §5.3. This Lyman-Werner background may therefore regulate the transition from very high mass primordial stars to the lower mass channel.

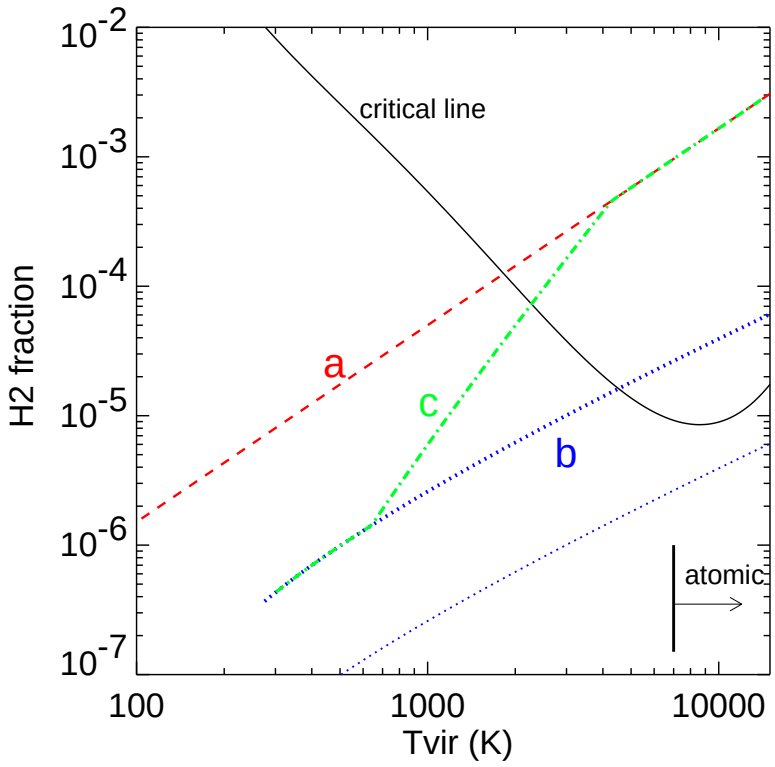

Figure 6.3 Schematic illustration of molecular hydrogen fraction as a function of virial temperature for virialized halos inside a cosmological simulation at $z = 17$. The solid line shows $f_{\mathrm{H}_2,c}$, the critical fraction required for efficient cooling. The dashed line (marked *a*) shows $f_{\mathrm{H}_2} \sim f_{\mathrm{H}_2,s}$ in the absence of radiative feedback (see Fig. 5.4). The thick (marked *b*) and thin dotted lines show f_{H_2} if self-shielding is neglected and $J_{\mathrm{LW},21} = 0.01$ or 0.1, respectively. The dot-dashed line (marked *c*) shows the same, but with self-shielding approximately included. Figure credit: Yoshida et al. 2003, ApJ, 592, 645.

6.2 THE X-RAY BACKGROUND: POSITIVE FEEDBACK

The radiative feedback on H_2 need not be only negative, however. In the dense interiors of gas clouds, the formation rate of H_2 could be accelerated through the production of free electrons by X-rays. This effect could counteract the destructive role of H_2 photo-dissociation. Unlike UV photons, X-rays can penetrate huge distances across the Universe, even at high redshifts. The comoving mean free path through the mean IGM density of an X-ray photon with energy E is:

$$\lambda_{\rm X} \approx 11\,\bar{x}_{\rm HI}^{1/3} \left(\frac{1+z}{10}\right)^{-2} \left(\frac{E}{300\,{\rm eV}}\right)^{3} {\rm comoving\,Mpc}; \qquad (6.12)$$

thus, photons with $E > 1.5[(1+z)/15]^{1/2}\bar{x}_{\rm HI}^{1/3}$ keV propagate an entire Hubble length before interacting with the IGM. Similarly, they can penetrate large columns of dense neutral gas inside of collapsed halos. Thus, any X-ray background would be pervasive at high redshifts.

X-rays interact with primordial gas, by either ionizing helium or hydrogen. The resulting free electron can gain a large kinetic energy (equal to the difference between the photon energy and the ionization potential), which it then deposits as a mixture of heat, collisional ionization, and collisional excitation. Typically, a fraction $f_i \sim x_{\rm HI}/3$ of the energy is deposited in ionizing other atoms. Thus, a 1 keV photon can result in ~ 25 free electrons. Because these free electrons catalyze H_2 formation, X-rays can exert a *positive* feedback on primordial star formation.

An X-ray background seems almost inevitable at high redshifts, with a number of possible sources: (1) Very massive Population III stars are hot enough for their blackbody spectra to extend into the soft X-ray regime. (2) Quasars or "mini-quasars" must begin to form at very high redshifts in order to produce the extremely luminous quasars seen at $z \sim 6$ and likely have nonthermal spectra extending to very high energies. (3) Supernova blastwaves may accelerate fast electrons, which can in turn scatter CMB photons to X-ray energies. The associated cooling rate of relativistic electrons increases dramatically with redshift since the CMB energy density scales as $u_{\rm CMB} \propto (1+z)^4$. (4) X-ray binaries, in which a massive black hole accretes gas from a companion, are often produced when a massive star explodes in a binary system; if massive stars are more abundant at high redshifts, then such binaries may be more common then. We will see later (§11.3.2) that these contributions to the X-ray background significantly affect the IGM temperature and ionization history, and they also present an important potential positive feedback mechanism for the first stars.

Simple scaling laws suggest, however, that this positive feedback will only overcome the negative Lyman-Werner feedback in unusual circumstances. Let us suppose that the electron fraction inside a cool cloud is in ionization equilibrium with an X-ray background. We will assume that the X-rays are sourced by the same population of galaxies as the ultraviolet background (though the sources themselves may differ, such as high mass stars and the X-ray binaries they form after dying).

For an X-ray background amplitude J_X, ionization equilibrium implies $n_e \propto (J_X n)^{1/2}$, where we have ignored the temperature dependence of the recombina-

tion coefficient. Equation (6.4) therefore yields

$$f_{\mathrm{H}_2,eq} \propto n^{1/2} J_X^{1/2} / J_{\mathrm{LW}}. \tag{6.13}$$

In other words, the equilibrium molecular fraction depends more weakly on the X-ray background than on the UV background. Assuming these are tied to the same underlying physical processes (i.e., are both ultimately driven by gas accretion onto halos and star or black hole formation), X-rays can only make a substantial difference when J_{LW} is still relatively modest. (Moreover, they only matter at all if the equilibrium electron fraction is larger than the value obtained from the usual chemistry described in §5.1.1.)

More detailed investigations have shown that if $J_X = \epsilon_X J_{\mathrm{LW}}$ at the H I ionization edge, X-rays exert mild positive feedback on dense gas clouds when $0.1 < \epsilon_X < 1$. At smaller fluxes, the X-rays are relatively unimportant. At larger fluxes, the heating generated by the X-rays counteracts the additional cooling, negating the boosted free electron fraction.[iii]

6.3 RADIATIVE FEEDBACK: MECHANICAL EFFECTS

As discussed in §5.2.2 that radiative feedback from the first stars may be crucial for choking off accretion and setting their final mass scale. But the high-energy photons responsible for that process likely reach much farther into the source halo and the surrounding IGM once the star enters the main sequence. The same processes mentioned previously can dramatically affect these regions and subsequent star formation in them, because the radiation can influence the motion of the surrounding gas more than gravity. In this section we will consider some of the relevant processes in more detail.

6.3.1 The First H II Regions: Photoevaporation

The most dramatic effects result from the high luminosity of ionizing photons produced by the first stars. We discussed briefly in §5.2.2 how ionization fronts can have powerful effects on gas dynamics – these effects extend far beyond the star once it enters the main sequence. For example, consider an ionizing front expanding inside a gravitationally-bound halo, where the baryon density declines with radius. For pedagogical purposes, we adopt a simple density profile:

$$n(r) = \begin{cases} n_c & r < r_c, \\ n_c (r/r_c)^{-w} & r \geq r_c, \end{cases} \tag{6.14}$$

where w is a power-law index that encapsulates the steepness of the profile and n_c and r_c are a core density and radius, respectively. Numerical simulations show that primordial gas clouds have $w \sim 2$–2.2 (see also §5.2.1).

The properties of the ionization front can be characterized with reference to the *Strömgren radius* R_s, the outer boundary around the source out to which the total

[iii] A similar negative feedback effect is at work even at lower X-ray fluxes in the diffuse IGM, where the X-ray heating creates an "entropy floor" that prevents collapse onto virialized objects. We discuss this effect in detail in §8.10.

rate of recombinations are equal to the total rate of ionizations. For a star producing ionizing photons at a rate $\dot{N}_i$ in a constant density medium, this radius is

$$R_s = \left(\frac{3\dot{N}_i}{4\pi n^2 \alpha_B}\right)^{1/3} \approx 150 \left(\frac{\dot{N}_i}{10^{50}\ \mathrm{s}^{-1}}\right)^{1/3} \left(\frac{n}{\mathrm{cm}^{-3}}\right)^{-2/3} \mathrm{pc}, \qquad (6.15)$$

where we have evaluated the recombination coefficient $\alpha_B = 2.6\times10^{-13}\,\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$ at $\sim 10^4$ K. If the H II region reaches this size (or, alternatively, if for a fixed radius the density exceeds an equivalent threshold value), then the ionizing photons themselves are consumed within mostly-ionized gas. Before this time, the front was only limited by the rate at which photons could ionize the medium. During this fast expansion phase, we refer to the ionization front as *R-type*. However, once this expansion velocity slows down to near the sound speed, the gas will be able to react to its new thermodynamic properties. The Strömgren radius provides a simple way to estimate when this transition occurs, because at that point the expansion will have nearly zero velocity. In more detail, the ionization front slows to become *D-type* when its expansion speed falls to roughly twice the isothermal sound speed of the ionized medium, $2c_i$. At this point, the increased temperature (and hence pressure) within the front drives a shock into the surrounding medium. The front then propagates outward at roughly the speed of sound. The H II region can therefore only expand through hydrodynamic processes and the ionization front is said to be *trapped*. The average density of virialized (uncooled) gas inside dark matter halos at redshift z is $\sim 1\,\mathrm{cm}^{-3}[(1+z)/30]^3$, independent of halo mass.

In the density profile given by equation (6.14), the Strömgren radius is $R_w \equiv g(w)R_s$, where R_s is evaluated using the core density and

$$g(w) = \begin{cases} \left[\frac{3-2w}{3} + \frac{2w}{3}\left(\frac{r_c}{R_s}\right)^3\right]^{1/(3-2w)} \left(\frac{R_s}{r_c}\right)^{2w/(3-2w)}, & w \neq 3/2 \\ \left(\frac{r_c}{R_s}\right) \exp\left\{\frac{1}{3}\left[\left(\frac{R_s}{r_c}\right)^3 - 1\right]\right\}. & w = 3/2. \end{cases} \qquad (6.16)$$

The front's speed will depend on how far it extends: it can *accelerate* at $r > r_c$ if the density profile is steep enough. In particular, if $w > 3/2$, the recombination rate ($\propto n^2$) does not appreciably increase as the front's radius grows, allowing the front itself to escape to infinity. To see this, it is straightforward to estimate the velocity at which the ionization front expands before the R_w limit is met:

$$U_{\mathrm{if}} = \frac{U_c}{(R_s/r_c)^3 - 1} u(w), \qquad (6.17)$$

where U_c is the typical speed within the core and

$$u(w) = \begin{cases} \left(\frac{r_c}{R}\right)^{2-w}\left[\left(\frac{R_s}{r_c}\right)^3 + \frac{2w}{3-2w} - \frac{3(R/r_c)^{3-2w}}{(3-2w)}\right] & w \neq 3/2 \\ \left(\frac{r_c}{R}\right)^{1/2}\left[\left(\frac{R_s}{r_c}\right)^3 - 1 - 3\ln(R/r_c)\right] & w = 3/2. \end{cases} \qquad (6.18)$$

The ionization front speed U_{if} will remain R-type all the way to infinity if

$$w > w_{\mathrm{trap}} = \frac{3}{2}\left[1 - \left(\frac{r_c}{R_s}\right)^3\right]^{-1}, \qquad (6.19)$$

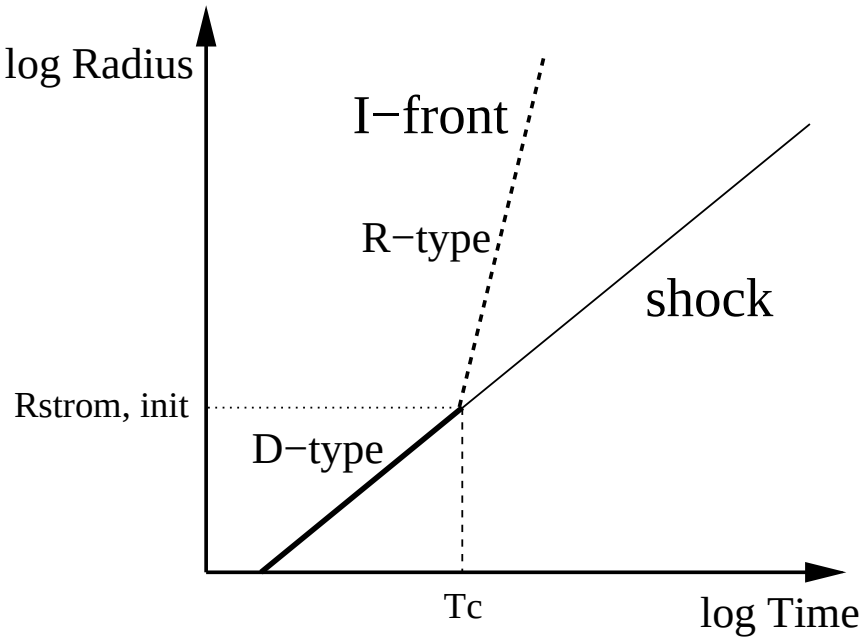

Figure 6.4 Cartoon of an ionization front propagating through a cosmological halo. After an initial R-type phase (not shown), recombinations in the high-density core trap the front, making a D-type front in which a shock leads the ionization front. As the density falls, the recombination rate also falls, eventually freeing the front to expand much faster than the sound speed. The shock is left behind and lags the front, often transforming into a simple pressure wave. Figure credit: Yoshida et al. 2007, ApJ, 663, 687.

or $w > 3/2$ for ionization fronts able to reach well outside the core before striking the Strömgren limit.

The shock front will shift to D-type, driving a shock into the surrounding gas, if $w < w_{\rm trap}$. This allows the ionization front to grow (slowly), even though it has nominally reached its Strömgren limit, because the hydrodynamic motions of the gas decrease the average density behind the shock. In a typical halo, the density profile steepens as one moves outward, usually with $w > w_{\rm trap}$ in the outskirts. Therefore, the front will eventually reach a point where it is no longer trapped. At this time it will revert to R-type and expand rapidly, with no immediate hydrodynamic effect on gas outside of the H II region itself. Numerical simulations show that this point is well approximated by the Strömgren radius of the initial density profile, using equation (6.15) with the average density inside R_s. Figure 6.4 shows a cartoon of this evolution.

However, within the H II region, the gas rapidly accelerates outward. The temperature structure of the cloud is set by photoheating: each ionization leaves the residual electron with some extra energy that depends upon the spectrum of the ionizing source (see §8.10 for more details on this process), typically with $T \sim 10^4$ K.

The pressure profile will be set by the density profile – which also has not had time to adjust to its new state. A strong pressure gradient therefore develops, producing an acceleration

$$a = \frac{1}{\rho}\frac{dp}{dr} \sim \frac{wc_i^2}{r}, \tag{6.20}$$

which is strongest in the center of the halo. A pressure wave therefore develops, pushing the gas ahead of it out of the halo – this regime is often referred to as the *champagne phase*. Behind the wave, the gas will have roughly constant density and hence reach pressure equilibrium; ahead of it the gas will still be in its original configuration.

The characteristic speed of this wave is a few times the sound speed of the ionized gas, $c_i^{\rm ion} \sim \sqrt{kT/m_p} \sim 10(T/10^4\ {\rm K})^{1/2}$km s^{-1}. In comparison, the escape speed from a dark matter halo is roughly

$$v_{\rm esc}(M) \approx \sqrt{2}V_c(r_{\rm vir}) = 24.0\left[\frac{\Omega_m}{\Omega_m(z)}\,\frac{\Delta_c}{18\pi^2}\right]^{1/6}\left(\frac{M}{10^8 M_\odot}\right)^{1/3}\left(\frac{1+z}{10}\right)^{1/2}\ {\rm km\ s^{-1}}\,. \tag{6.21}$$

where we have used equation (3.32) and assumed an isothermal density profile truncated at the virial radius $r_{\rm vir}$, for simplicity. Thus, the gas inside the H II region becomes strongly unbound and flows out of its host for halos of a sufficiently low mass, M. The ionization front will only slow down when it reaches a region with a shallower density gradient in the IGM, allowing it to return to the Strömgren limit. But by this point the bound gas had already escaped.

Numerical simulations of this *photoevaporation* process show that, in the limit of a smooth, spherical halo, the radiation pressure from a single Population III.1 very massive star can evacuate the gas from an entire halo of mass $\sim 10^6\ M_\odot$. Figure 6.5 shows an example from a detailed numerical calculation. The simulation takes a single $200\ M_\odot$ star at $z = 18.2$ in a halo of total mass $7\times10^5\ M_\odot$. Clockwise from top left, the panels show the ionized fraction, the temperature, the (outward) velocity, and the density profile. Each panel shows snapshots at 63, 82, 95, 127, 317, and 2200 kyr (left to right in all panels except bottom left, where they are top to bottom). In the last panel, the dashed line shows the density required to enforce the Strömgren criterion in equation (6.15); if the density exceeds this value, the ionization front will be limited by recombinations and be D-type. Clearly, the high core density will trap the front, from which it will emerge when it reaches ~ 1 pc. In the upper panels, the large jump in the ionization front location from 82–95 kyr involves the transition from D-type to R-type. The large outward gas velocities, at 2–3 times the sound speed of the gas are nearly ~ 10 times as large as the escape speed from the minihalo (2–3km s^{-1}).

Thus, the first stars have nearly emptied their halos of gas, decreasing the baryon fraction to just a few percent. However, the picture is less clear if the gas filling the source halo is clumpy, with other collapsing cores (or in nearby halos also en route to forming their own stars). If these neighboring clumps have modest densities, they too will be completely evaporated. However, if their central densities are sufficiently high, $n > 2000\ {\rm cm}^{-3}$, the core will remain neutral via self-shielding, the radiation will have little effect, and the collapse will continue until new stars are

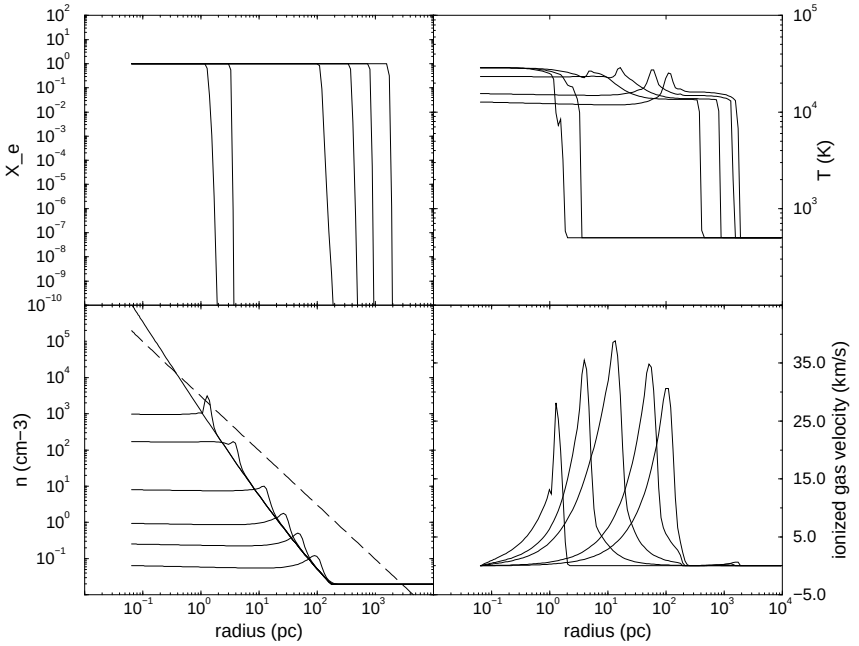

Figure 6.5 Evolution of a cosmological halo as an ionization front propagates through it. The simulation takes a single 200 $M_\odot$ star at $z = 18.2$ in a halo of total mass 7×10^5 $M_\odot$. Clockwise from top left, the panels show the ionized fraction, the temperature, the (outward) velocity, and the density profile. Each panel shows snapshots at 63, 82, 95, 127, 317, and 2200 kyr (left to right in all panels except bottom left, where they are top to bottom). In the bottom left, the dashed line shows the density required to trap the ionization front. Figure credit: Whalen et al. 2004, ApJ, 610, 14.

formed. The passage of a through surviving cores shock may actually aid collapse and encourage further star formation. Whether more than one star can form in a low-mass halo (or in a halo with nearby neighbors) thus crucially depends on the degree of synchronization of clump formation.

As an example of the complex implications of the photoevaporative flow, consider the shocked gas that lies ahead of the front during its D-type phase. This shocked region is partially ionized by high-energy photons and so its ionized fraction is typically appreciable ($> 10^{-3}$). The extra free electrons catalyze the formation of H_2, potentially making self-shielding effective. The cooling induced by H_2 can trigger fast thin shell instabilities that develop into new star-forming clumps.

The long-term effects of this radiation pressure are also not obvious and depend on the details of the halo's neighborhood. Although the gas has very high velocity as it leaves the halo, it can still be reincorporated into the halo (or into one of its nearby neighbors) through hierarchical structure formation. But numerical simulations show that the fallback can take ~ 100 Myr, a substantial fraction of the age of the Universe at these high redshifts. This could lead to a long delay in later star formation or accretion onto any remnant black holes. The pre-ionization would also change the mode of any future star formation to Population III.2 stars, possibly with a somewhat lower mass scale than the first generation of Population III.1 stars.

6.3.2 Radiation Pressure From Lyman-α Photons

Interestingly, the radiation can also exert a substantial force on the neutral gas surrounding the H II region. The Lyman-α photons generated primarily by recombinations within the H II regions scatter off the outside gas, imparting their net outward momentum and driving the gas away from the central source. We can gauge the possible dynamical effect of these photons by comparing the gravitational binding energy, $E_B \sim (\Omega_b/\Omega_m)GM^2/r_{\rm vir}$, with the energy in the radiation field, $E_\alpha = L_\alpha \times t_{\rm trap}$, where L_α is the line luminosity of H II region and $t_{\rm trap}$ is the typical timescale over which Lyman-α photons are trapped inside the cloud. Numerical calculations of line transfer suggest that $t_{\rm trap} \sim 15 t_{\rm light}$, where $t_{\rm light} = r_{\rm vir}/c$ is the light travel time across the halo (see more discussion of this complex problem in §10.1.1). The condition $E_\alpha > E_B$ requires

$$L_\alpha > L_{\alpha,{\rm crit}} \sim 10^{40} \left(\frac{M}{10^6\ M_\odot}\right)^{4/3} \left(\frac{1+z}{30}\right)^2 \left(\frac{15 t_{\rm light}}{t_{\rm trap}}\right) {\rm erg\ s^{-1}}. \quad (6.22)$$

Note that $\sim 2/3$ of recombinations produce a Lyman-α photon, so this translates to a direct constraint on the ionizing luminosity; the fiducial luminosity shown here corresponds to only $\sim 500\ M_\odot$ (per $M \sim 10^6 M_\odot$) in very massive Population III.1 stars, assuming that the H II region reaches its Strömgren limit.

For a nearly isotropic radiation field (valid in this case because of the large number of scatterings each Lyman-α photon experiences), the acceleration induced by Lyman-α radiation pressure may be written as

$$a_{{\rm Ly}\alpha} = \frac{1}{3\rho}\frac{dU_\alpha}{dr}, \quad (6.23)$$

where $\rho = m_H n$, and U_α is the energy density of the Lyman-α photons. If the gas was optically-thin, then U_α would have been $L_\alpha/(4\pi r^2 c)$, but the scattering process traps Lyman-α photons near the source and steepens the $1/r^2$ scaling. The total impulse $a_{\mathrm{Ly}\alpha}\Delta t$ therefore depends on the total Lyman-α fluence of the source, which in turn is dictated by the number of ionizing photons produced by the stars.

The simple solution described in §10.4.1, for scattering around a point source in a uniform IGM expanding at the Hubble flow, has $U \propto r^{-2/3}$ at moderate distances from the source. In reality, the H II region surrounding the central star, the infall region surrounding the halo, and the details of Lyman-α scattering must be taken into account in a realistic calculation, but this simple solution provides a reasonable gauge of the importance of Lyman-α radiation pressure. Assuming very massive Population III.1 stars, the corresponding final velocity of an atom at a distance r from the central source is

$$v_\alpha \sim 6 \left(\frac{1\ \mathrm{kpc}}{r}\right)^{10/3} \left(\frac{15}{1+z}\right)^3 \left(\frac{f_\star}{10^{-3}} \frac{M}{10^6\ M_\odot}\right) \mathrm{km\ s^{-1}}. \quad (6.24)$$

While the final velocity is small, the escape speed at the virial radius $r_{\mathrm{vir}} = 0.2$ kpc of a $10^6\ M_\odot$ halo at $z = 14$ is $\sim$ 6km s^{-1}. Thus, Lyman-α scattering through the neutral gas *outside* of any H II region can eject the gas from the vicinity of the source halo, also slowing down accretion.

This same effect can also operate in larger galaxies later on in the history of structure formation, many of which are observed to have substantial Lyman-α fluxes. However, numerical simulations show that the effects are modest unless the galaxy also drives a wind that creates a neutral "supershell" that can multiply the radiation force through repeated scatterings. This is largely because these galaxies are able to ionize such a large region around them that the near region, where the force is strongest, is still ionized and cannot trap the photons.

6.4 WINDS AND MECHANICAL FEEDBACK

6.4.1 Star Formation and Wind Energetics

As stars live and die, they inject large amounts of energy into their surroundings, through a number of channels. First, while they are luminous, their radiation couples to the interstellar medium as UV photons scatter off of dust grains (which are usually coupled to the neutral or ionized gas through collisions and magnetic fields). Just as in the Lyman-α scattering case described above, the pressure of the radiation field can therefore eject gas from the galaxy.

Second, in the late stages of stellar evolution, many stars drive powerful winds into the ISM. Finally, supernova explosions when stars die inject $E \sim 10^{51}$ ergs of energy into the ISM, typically accelerating $\sim 10\ M_\odot$ of material per explosion to $\sim 3 \times 10^3$km s^{-1}. The energy and momentum flux from these mechanical interactions can also unbind the gas from the host halo. By removing gas from the galaxy, these mechanisms choke off the fuel supply for further star formation and may ultimately be responsible for regulating the pace of star formation over time.

A clear understanding of the role of feedback is therefore essential to understand not only the first galaxies but their more massive descendants later on.

We begin with some plausibility arguments showing that winds are likely to be important for the small galaxies most common at high redshifts. We first ask the question of how much star formation is necessary to unbind the gas inside of a virialized halo. The total binding energy of a halo with mass M is given by equation (3.34), but for the *gas* we must multiply this by the mass fraction in gas ($f_g \sim \Omega_b/\Omega_m$). Moreover, we have already seen that in order to form stars the gas must collapse to high densities. To describe this simply, we assume that the gas is confined to a region $< \lambda r_{\rm vir}$ (see §9.5.3 below). Thus $E_{\rm b,g} \sim (f_g/\lambda)GM^2/r_{\rm vir}$. Meanwhile, the energy injected by supernovae is $E_{\rm SN} \sim f_\star f_g M \omega_{\rm SN}$, where $f_\star$ is the fraction of gas that is turned into stars and $\omega_{\rm SN}$ is the supernova energy input per unit mass of star formation. Typical supernova models and Population II IMFs yield $\omega_{\rm SN} \sim 10^{49}$ erg $M_\odot^{-1}$. However, we expect that some fraction of this energy will be radiated away as the hot, dense supernova remnant plows through the galaxy into the IGM around it. We assume that a fraction ξ of the total energy is available for mechanically removing gas from the galaxy. Then the energy input by supernovae exceeds the binding energy of the gas if

$$f_\star > f_{\star,E} \sim 0.01 \left(\frac{0.05}{\xi\lambda}\right) \left(\frac{M}{10^8\,M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \left(\frac{\omega_{\rm SN}}{10^{49}\,{\rm erg}\,M_\odot^{-1}}\right)^{-1}. \tag{6.25}$$

Even if the supernova remnants do lose their thermal energy, they will still inject a great deal of momentum into the ISM. If this momentum is large enough, it can carry the gas outside of the halo without the "push" from the thermal energy inside each remnant (i.e., feedback can be much more effective than suggested by Eq. (6.25) if $\xi \ll 1$). The rate at which momentum is injected by supernovae, $dP_{\rm SN}/dt$, is

$$\frac{dP_{\rm SN}}{dt} \sim 2 \times 10^{33} \left(\frac{\omega'_{\rm SN}}{300{\rm km\ s^{-1}}}\right) \left(\frac{\dot{M}_\star}{M_\odot\ {\rm yr}^{-1}}\right)\ {\rm g\ cm\ s^{-2}}, \tag{6.26}$$

where $\omega'_{\rm SN}$ is the rate of momentum injection from supernovae per unit mass of stars; the fiducial value takes one explosion per $100\,M_\odot$ of stars, each accelerating $\sim 10\,M_\odot$ of material to $\sim 3 \times 10^3$km s^{-1}.

Meanwhile, the rate at which stellar radiation injects momentum is $dP_{\rm rad}/dt \sim L/c$, where L is the stellar luminosity that couples to the interstellar medium (ISM) gas.[iv] We write it in terms of the rest energy as $L = \epsilon \dot{M}_\star c^2$, where $\epsilon_3 \equiv (\epsilon/10^{-3}) \sim 1$ for typical IMFs. Then $dP_{\rm rad}/dt \sim \epsilon_3 dP_{\rm SN}/dt$, indicating that both sources of momentum are likely important in launching winds.

The acceleration equation for a parcel of gas with velocity v and position r is

$$\frac{dv}{dt} = -\frac{GM(r)}{r^2} + \frac{L}{cM_g(r)}, \tag{6.27}$$

[iv] We assume here that the gas is optically thick to the radiation, so that it efficiently absorbs the momentum flux. If $\tau < 1$ (for example, because the metallicity is small and dust is rare), then the momentum injection rate from radiation decreases $\propto \tau$.

where $M(r)$ is the halo mass and $M_g(r)$ is the gas mass enclosed within a radius r. For a simple estimate, let us assume that the halo is a singular isothermal sphere, with $M(r) = 2\sigma^2 r/G$ where σ is the velocity dispersion, and that the gas traces the dark matter. Then we can rewrite equation (6.27) as

$$\frac{dv}{dt} = \frac{2\sigma^2}{r}\left(\frac{L}{L_M} - 1\right), \tag{6.28}$$

where

$$L_M = \frac{4 f_g c}{G}\sigma^4. \tag{6.29}$$

Clearly, L_M represents the minimum luminosity for the net force on the gas parcel to act outward, and hence it is the minimum luminosity required in order to launch a wind. If we further assume a constant star formation efficiency $f_\star$ to convert gas into stars over a dynamical time $t_{\rm dyn} \sim r_{\rm vir}/\sigma$, this minimum luminosity translates into a minimum star formation efficiency $f_{\star,p}$:

$$f_\star > f_{\star,p} \sim 0.1(\omega'_{\rm SN,300} + \epsilon_3)\left(\frac{M}{10^8\ M_\odot}\right)^{1/3}\left(\frac{1+z}{10}\right)^{1/2}, \tag{6.30}$$

where $\omega'_{\rm SN,300} = \omega'_{\rm SN}/(300{\rm km\ s^{-1}})$ and the $(\omega'_{\rm SN,300} + \epsilon_3)$ factor accounts for both supernovae and radiation.

Comparing equations (6.25) and (6.30), it is clear that for the small halos in which the first stars form, the energy reservoir is likely much more important than the raw momentum, provided that it is not lost through radiative cooling. It is also clear that the required star formation rate in these halos is very small: this is fundamentally because the energy available in stars scales with M (assuming a constant $f_\star$), while the binding energy scales as M^2.

However, at higher masses the excess energy becomes less important, with the momentum injection condition becoming more important when

$$M > M_p \sim 10^{11}(\omega'_{\rm SN,300} + \epsilon_3)^3\left(\frac{0.05}{\lambda\xi}\right)^3\left(\frac{1+z}{10}\right)^{3/2} M_\odot. \tag{6.31}$$

Nevertheless, in order for the momentum to lift gas out of the halo, star formation must proceed very quickly - turning a substantial fraction of the gas into stars over a single dynamical time. Such rates do appear in rapidly star-forming galaxies at lower redshifts, but those systems are relatively rare.

These two types of winds, *energy-driven* and *momentum-driven* are likely to have very different characters. The condition that $E_{\rm SN} > E_{\rm b,g}$ does not place any restrictions on the rate at which *mass* is ejected from the galaxy; in fact, numerical simulations of star-forming disk galaxies typically show that the energy is "blown-out" along low-column density channels (perpendicular to the disk), carrying away only a fraction of the galaxy's mass. On the other hand, the momentum must keep its direction and sweep any gas it encounters as it propagates outwards, carrying with it a significant fraction of the galaxy's gas. The asymptotic velocity v_∞ of a momentum-driven wind is typically just a few times the escape speed of the halo.[v]

[v]This can be seen, for example, by integrating equation (6.28) under the assumption that L is constant to obtain $v(r)$. Writing $(dv/dt) = (dr/dt)/(dv/dr) = v(dv/dr) = (d(\frac{1}{2}v^2)/dr$ and integrating both sides over $r > r_0$, yields $v(r) = 2\sigma \times [(L/L_M - 1)\ln(r/r_0) + v^2(r_0)/4\sigma^2]^{1/2}$.

Momentum conservation then demands that the mass loss rate in the wind is

$$\dot{M}_w = \frac{dp/dt}{v_\infty} \sim \dot{M}_\star \left(\frac{300(\omega'_{\rm SN,300} + \epsilon_3)\text{km s}^{-1}}{v_\infty} \right), \tag{6.32}$$

comparable to the star formation rate for reasonably large halos.

6.4.2 Expanding Blastwaves: Simple Solutions

In order to better understand the dynamics of these winds, we will review here some simple models for expanding blastwaves following point explosions. Although oversimplified, these analytic scalings provide useful insight into the more complex problem of winds inside and outside galaxies.

First consider a point explosion with energy E in a static, cold (or pressureless) medium of mass density ρ. The explosion drives a shock into the surrounding gas. Simple dimensional arguments show that the shock radius must depend on ρ, E, and time t in the form

$$R_{\rm sh} = K_{\rm STV}(Et^2/\rho)^{1/5}, \tag{6.33}$$

where $K_{\rm STV}$ is a constant.

It is easy to show from energy conservation that $K_{\rm ST} \sim 1$. The total mass that has been swept up by the shock is $\sim \frac{4\pi}{3}\rho R_{\rm sh}^3$. Because a supersonic shock forms in the ambient medium, the post-shock gas velocity must be subsonic in the frame of the shock. Thus, most of the bulk velocity of the material will be from the shock itself, and the net fluid speed is $\sim U_{\rm sh} = \frac{2}{5}R_{\rm sh}/t$. The kinetic energy of the swept-up material is therefore $\sim \frac{4\pi}{3}\rho R_{\rm sh}^3 \times \frac{1}{2}U_{\rm sh}^2 \sim 0.3R_{\rm sh}^5/t^2$. There is also, of course, the thermal energy stored in the hot gas behind the shock, but this is typically comparable to the kinetic energy of the shock. Energy conservation implies $E = \kappa\rho R_{\rm sh}^5/t^2$, where κ is a constant of order unity that accounts for summing the kinetic and internal energies. By comparison to equation (6.33), we see that $K_{\rm STV} = \kappa^{-1/5}$, which we expect to be very close to unity. In fact, this problem can be solved analytically, giving the exact value of $K_{\rm STV} = 1.17$ for a pure monatomic gas with an adiabatic index $\gamma = 5/3$. The solution is known as a *Sedov-Taylor-von Neumann blastwave*, after the three physicists to derive it independently at the dawn of the nuclear weapons age. Since there is no characteristic timescale or lengthscale in the setup of a point explosion, the hydrodynamic equations admit a *self-similar* solution in which the hydrodynamic variables of the gas (pressure, density, and velocity) depend only on the combination $r/R_{\rm sh}(t)$ instead of depending separately on r and t.

The Sedov-Taylor-von Neumann solution imposes three restrictions on the blastwave. First, it requires that the mass of the material behind the shock is much greater than the explosion ejecta. In an earlier phase, the ejecta expands ballistically encountering negligible resistance by the ambient medium. Second, it assumes a strong shock, or that the ejecta velocity greatly exceed the sound speed of the ambient medium. Finally, it assumes that all the explosion energy is contained either in the kinetic energy or thermal energy of the shocked gas. In fact, the strong shock jump conditions require that the density just behind the shock is

$(\gamma+1)/(\gamma-1) = 4$ times that of the ambient medium, decreasing rapidly inwards. This overdense shell will cool radiatively; once a substantial fraction of the energy has been lost, the energy conservation condition no longer applies and the character of the solution changes. In particular, as the gas in the shell cools its density must increase to maintain pressure equilibrium with the interior of the blastwave, and so a dense shell develops at the leading edge of the blastwave.

This second phase is known as a *pressure-driven snowplow*, because the low-density interior of the gas remains hot (and hence has a finite pressure p pushing outward on the shell). In this phase, the shell sweeps up gas as it expands, increasing its mass at a rate $\dot{M}_s = 4\pi R_{\rm sh}^2 \rho U_{\rm sh}$. Meanwhile, so long as the hot interior does not cool, the internal pressure obeys the adiabatic condition $pV^\gamma =$constant, pushing the shell outward with a force $4\pi R^2 p$. The equation of motion for the shell is then

$$\ddot{R}_{\rm sh} + \frac{3\dot{R}_{\rm sh}^2}{R_{\rm sh}} = \frac{3p_i}{\rho R_{\rm sh}} \left(\frac{R_i}{R_{\rm sh}}\right)^{3\gamma}, \tag{6.34}$$

where p_i is the internal pressure as this phase begins when $R_{\rm sh} = R_i$. For $p_i \neq 0$ and $\gamma = 5/3$, this equation requires that $R_{\rm sh} \propto t^{2/7}$, slightly shower than in the "adiabatic" Sedov-Taylor-von Neumann phase.

The pressure-driven snowplow phase ends when either one of two conditions is fulfilled. First, if the hot bubble interior can cool radiatively, it loses the pressure support. Second, if the interior pressure approaches the pressure of the ambient medium, there will be no net driving force. In either case, $p_i \rightarrow 0$ in equation (6.34). In that case, $R_{\rm sh} \propto t^{1/4}$, which follows strictly from momentum conservation, $\frac{4\pi}{3}\rho R_{\rm sh}^3 (dR_{\rm sh}/dt) = const.$ This final phase is therefore known as a *momentum-conserving snowplow*. Obviously, it is the proper solution for the momentum-driven winds described in the previous section.

So far we have assumed that the blastwave propagates into a uniform medium. While this describes the ISM of normal galaxies reasonably well, the gas making the first stars had not settled into disk-like configurations; instead, these stars were surrounded by uniform density cores inside roughly power law envelopes, with $\rho \approx \rho_0 (R_0/R)^\alpha$. Dimensional arguments similar to those above then show that

$$R_{\rm sh} = K_{\rm iso} \left(\frac{Et^2}{\rho_0 R_0^\alpha}\right)^{1/(5-\alpha)}. \tag{6.35}$$

In particular, for an isothermal density profile $\alpha = 2$, close to the envelopes of the first stars, $R_{\rm sh} \propto t^{2/3}$. The blastwave propagates faster in this case because the declining ambient density presents considerably less drag.

Similarly, it is straightforward to modify the equation of motion for the snowplow shell: since $\dot{M}_s = 4\pi R^2 \rho(R) U_{\rm sh}$, the momentum equation reads

$$\ddot{R}_{\rm sh} + \frac{(3-\alpha)\dot{R}_{\rm sh}^2}{R_{\rm sh}} = \frac{(3-\alpha)p_i}{\rho R_{\rm sh}} \left(\frac{R_i}{R_{\rm sh}}\right)^{3\gamma}. \tag{6.36}$$

For $\gamma = 5/3$, this equation admits the solution $R_{\rm sh} \propto t^{2/(7-\alpha)}$ when the pressure is important, and $R_{\rm sh} \propto t^{1/(4-\alpha)}$ during the momentum-conserving snowplow phase. Again specializing to an isothermal density profile, $R_{\rm sh} \propto t^{2/5}$ and $R_{\rm sh} \propto t^{1/2}$ in these two phases.

6.4.3 Supernovae in the First Star-Forming Halos

The first supernovae occur in the halos described in §5. Although the basic properties of these halos are well-understood, the mass spectrum of stars and efficiency of star formation are highly uncertain as they depend on the complex fragmentation process, the degree of synchronization of the resulting protostellar clumps, and the dynamical impact of the surrounding H II region.

For these reasons, the overall impact of the first supernovae on their host halos is difficult to assess. Nevertheless, numerical simulations have begun to explore these events and their implications for subsequent star formation, at least in some simple cases. Figure 6.6 provides an example, showing a simulated supernova explosion of a 200 $M_\odot$ star at $z \approx 20$ in a halo with $M = 5 \times 10^5\ M_\odot$ and $r_{\rm vir} \approx 100$ pc (the box measures $150h^{-1}$ comoving kpc across). We will examine this result in some detail because it illustrates much of the important physics of high-redshift supernovae.

The grayscale shows the gas temperature. The large, roughly spherical region filling most of the box in all four panels is the H II region; its internal structure is a result of the filamentary cosmic web surrounding the halo. By the star's death (2 Myr after its formation), it has photoevaporated the gas inside $\sim r_{\rm vir}/2$, reducing its density to $n \sim 0.5\ {\rm cm}^{-3}$. Meanwhile, the escaping photons ionize a large region around the halo, initially heating it through inverse Compton scattering and causing pressure-driven expansion of the remnant into the low-density, cool IGM surrounding it.

The supernova then expands into this ionized environment. Figure 6.6 shows snapshots of its evolution, while Figure 6.7 presents the evolutionary phases of its (spherically-averaged) radius. The four major phases of the expansion are marked. The explosion here, which is assumed to completely blow apart the star via a pair-instability supernova, carries a substantial mass $M_{\rm ej}$ in ejecta. Until the swept-up mass dominates the explosion, it expands freely ('FE' in Figure 6.7). The simulation does not follow this short phase explicitly; instead it initializes the calculation at the end of this phase.

After that point, the Sedov-Taylor-von Neumann phase begins (marked 'ST'). The blastwave initially propagates through a roughly constant density interior (the remnant gas after photoevaporation), so $R_{\rm sh} \propto t^{2/5}$. Once the remnant reaches $\sim r_{\rm vir}/2$ (at $t \sim 10^5$ yr), it catches up to the photoevaporation shock, and the character of its surroundings change. However, at just about this time the gas in the dense shell accumulating behind the shock is able to cool. Several processes allow cooling: atomic (and molecular) line radiation, bremsstrahlung, and inverse Compton scattering of CMB photons. Ignoring any possible chemical enrichment from the supernova itself, the atomic cooling rates are shown in Figure **??**. Because these are driven by collisions, their rate scales as n^2. This mechanism is particularly important in dense gas, where it dominates over the other processes within the remnant shell. The cooling time is therefore $t_{\rm cool} \sim nkT/\Lambda \sim 10^5$ yr, where the initial temperature is $T \sim 10^6$ K.

Thus, at about the same time the remnant reaches the photoevaporation shock, the shell gas cools and transitions to a pressure-driven snowplow solution ('PDS' in

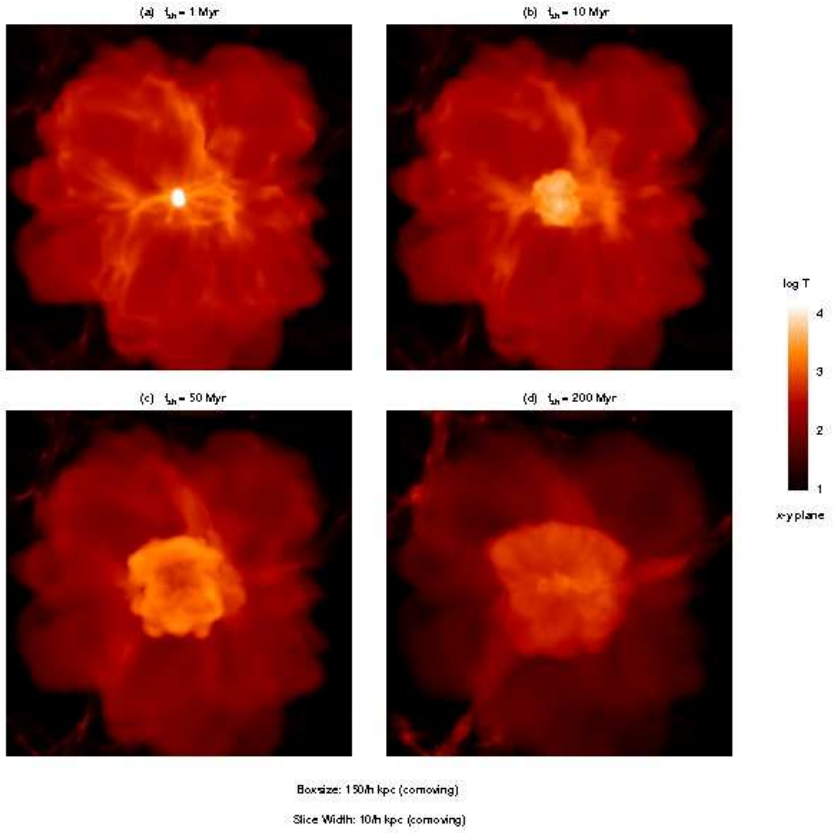

Figure 6.6 Temperature maps from a numerical simulation of a supernova explosion. The supernova of a 200 $M_\odot$ star is set off at $z \approx 20$ in a halo with $M = 5 \times 10^5\ M_\odot$ and $r_{\rm vir} \approx 100$ pc. The snapshots are 1, 10, 50, and 200 Myr after the explosion. In the first panel on the top left, the supernova is the central hot region; the star's H II region fills most of the box. The supernova remnant expands over the four panels, gradually becoming more anisotropic as it encompasses the filamentary structure surrounding the halo. Figure credit: Greif et al. 2007, ApJ, 670, 1.

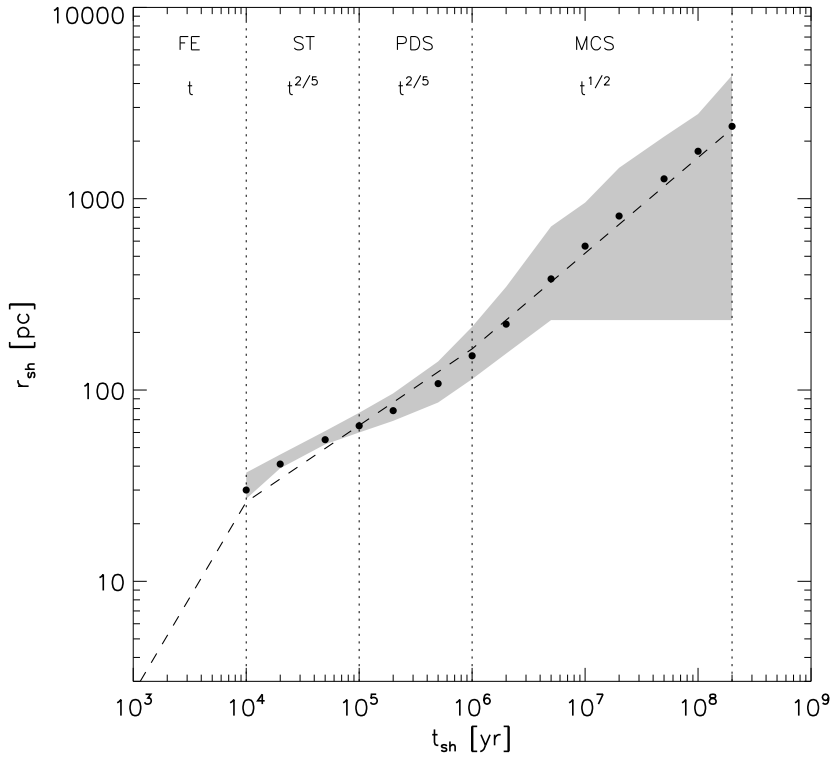

Figure 6.7 Evolution of the simulated supernova explosion described in Fig. 6.6. The black dots indicate the spherically-averaged mass-weighted shock radius, while the dashed line shows the analytic estimate using the models of §6.4.2. The different phases in the evolution of the remnant are labelled: 'FE' for free expansion (not resolved by the simulation), 'ST' for Sedov-Taylor-von Neumann phase, 'PDS' for pressure-driven snowplow, and 'MCS' for momentum-conserving snowplow. The shaded gray region shows the radial dispersion of the shock, which increases dramatically once the shock leaves its host halo. Figure credit: Greif et al. 2007, ApJ, 670, 1.

the figure). Now it propagates through the roughly isothermal sphere profile of the unperturbed halo, so $R_{\rm sh} \propto t^{2/5}$ – the same dependence as in the previous phase. This phase continues until either (1) the low-density interior gas is able to cool or (2) the postshock pressure reaches equilibrium with the ambient medium. At the very low densities characteristic of the remnant's interior, atomic cooling is inefficient. However, the cooling time due to inverse Compton scattering is independent of density (see Eq. 2.24),

$$t_{\rm cool} = 8 \left(\frac{20}{1+z} \right)^4 \ {\rm Myr}, \tag{6.37}$$

which puts an upper limit on the duration of this phase. But the postshock pressure reaches a value $p_{\rm sh} \sim \rho U_{\rm sh}^2 \sim p_{\rm HII}$, where $p_{\rm HII}$ is the pressure inside the H II region, after only $\sim 10^5$ yr. (This is easy to show using the analytic scalings of the previous section.)

Thus, the blastwave transitions to its final phase, the momentum-conserving snowplow (marked 'MCS' in Figure 6.7), at $\sim 10^5$ yr. At the beginning of this phase, the density profile is still roughly isothermal, so $R_{\rm sh} \propto t^{1/2}$; in the simulation $R_{\rm sh}$ maintains this scaling even after passing into the IGM (see §6.5.2 for a discussion of solutions in this limit).

The net effect of this single supernova is to completely disrupt the gas in the host halo, expelling much of it ($\sim 95\%$) and forcing the rest to high temperatures and low densities where star formation is inefficient. The lack of star formation will persist until the high-entropy gas can be reincorporated through hierarchical build-up of higher mass halos. Supernovae may therefore be efficient in quenching star formation within the first star-forming halos.

However, as in so many other aspects of feedback, there are a number of subtleties to this simple picture, some of which may actually *promote* further star formation. These include:

- First, the supernova itself is a source of heavy elements. As many of these elements are much more efficient low-temperature coolants than H or H_2, their presence could promote future star formation, particularly in combination with some of the mechanisms mentioned below. The primary uncertainty is the degree of mixing of the enriched material with the ambient medium, which is likely driven by instabilities in the shocked layers. We discuss the physics of this change in star formation in more detail in §**??**.

- Second, if the host halo remains largely neutral (for example, because the characteristic mass scale of Population III.1 stars is only $\sim 10\ M_\odot$, and they form in massive halos, so that the explosion energy is lower than the gravitational binding energy of the halo gas), the remnant will plow through much denser gas, even approaching $\sim 10^7$ cm^{-3}. Bremsstrahlung (free-free) cooling in such dense environments is extremely fast, and the supernova loses its thermal energy long before escaping the halo. Nevertheless, the impulse provided by the explosion can efficiently stir up the gas, possibly triggering further fragmentation as shells collide and most likely dispersing heavy elements throughout the halo.

- The blastwave itself may have very different effects on dense clumps (either inside the host galaxy or in nearby minihalos) than on the diffuse gas we have discussed. In particular, the shock compresses the gas, which increases the density, speeding up the later stages of collapse, provided that the gas can efficiently cool (which is a precondition for star formation). Furthermore, the ram pressure of the shock will likely not be able to move entire clumps along with the flow. The resulting configuration – a stream of moving fluid flowing by a stationary cloud – can be unstable to Kelvin-Helmholtz modes. If so, the resulting mixing may allow metals to penetrate the outer layers of the pristine minihalo gas, triggering a change in the mode of star formation.

- Finally, the dense shell that accumulates behind the leading shock can itself be unstable and fragment through gravitational or cooling instabilities into protostellar clumps. The condition for such fragmentation is similar to the classical Jeans instability: when the self-gravity of the shell operates faster than the restoring pressure forces in the shell, which implies that scales $> c_s/\sqrt{G\rho}$ collapse. For a given ambient density, the shell therefore eventually becomes unstable once the sound speed (or temperature) falls far enough, which of course requires efficient radiative cooling. In the case described in Figures 6.6 and 6.7, no fragmentation occurred because the low-density ambient medium both increased the dynamical time and inhibited molecule formation, maintaining relatively high temperatures. However, fragmentation can be much more efficient if the blastwave propagates through a denser neutral medium. In this case, the shell can trigger a second-generation of stars. Because these stars form out of ionized gas (either from a pre-existing H II region, or one produced by the passing shock), they will be similar to Population III.2 stars discussed in §5.3, with lower characteristic masses.

6.5 METAL ENRICHMENT AND THE TRANSITION TO POPULATION II STAR FORMATION

The very first stars formed under conditions that were much simpler than the highly complex birth places of stars in present-day molecular clouds. As soon as the first stars appeared, however, the situation became more complex due to their feedback on the environment. In particular, supernova explosions dispersed the heavy elements produced in the interiors of the first generation of stars into the surrounding gas. Atomic and molecular cooling became much more efficient after the addition of these metals.

Early metal enrichment and dispersal by the primordial supernovae described in the previous section, triggered a change in the fundamental mode of star formation, because heavy elements can radiatively cool the gas much more efficiently than H_2. To see this, consider a primordial cloud at the "loitering" phase with $n \sim 10^4$ cm^{-3} and $T \sim 200$ K. At this point, radiative cooling by H_2 becomes inefficient, so the gas contracts only slowly, and fragmentation is suppressed at least until an accretion disk forms around the first protostar. This is why the characteristic mass

of Population III.2 stars may be as high as $\sim 100\ M_\odot$ (see §5.1).

Now let us imagine that the gas instead has a small fraction of metals; if these elements can efficiently cool the gas from this thermodynamic state, they will induce further fragmentation to smaller mass scales. We will use the common notation [X/H]$= \log_{10}(N_X/N_H) - \log_{10}(N_X/N_H)_\odot$ to describe the abundance of species X. Detailed calculations show that carbon and oxygen are the most important elements at the relevant temperature and density, at least for atomic cooling. Carbon is likely to be singly-ionized C II, because the Universe is transparent to photons above the ionization potential of C I (11.26 eV), though it is sufficiently close to the Lyman-Werner bands that it may suffer some self-shielding by H_2 in very dense clumps. Oxygen, on the other hand, has an ionization potential very near H I (13.6eV) and so it will remain neutral. Let us write $\Lambda_X(n,T)$ for the radiative cooling rate from species X and $\Lambda_{\rm tot}$ for the total rate. For these two species and at the relevant temperatures and densities, the cooling is dominated by fine-structure lines of O I (wavelength of 63.1μm) and C II (157.7μm).

Fragmentation requires that the cooling time, $t_{\rm cool} = 1.5nk_BT/\Lambda_{\rm tot}$ be smaller than the free-fall time in the gas, $t_{\rm ff} \approx 1/\sqrt{G\rho}$. For a given species, this defines a *critical metallicity* $[{\rm X/O}]_{\rm crit}$ above which radiative cooling suffices to induce fragmentation. Detailed calculations of the fine structure transitions in these elements yield $[{\rm O/H}]_{\rm crit} \approx -3.0$ and $[{\rm C/H}]_{\rm crit} \approx -3.5$, with a factor of ~ 2 uncertainty owing to uncertainties in the thermodynamic state of the loitering phase.

The above considerations include only gas-phase cooling; many of the ISM metals at low redshifts are contained in dust grains, which can also aid cooling due to both thermal emission and H_2 formation (which can occur very efficiently on the surface of dust grains, since hydrogen atoms are trapped in close proximity). Locally, dust formation is generally attributed to winds in asymptotic giant-branch stars. At high redshifts, dust may be mainly produced in the metal-rich ejecta of supernovae themselves. The dust formed inside supernova ejecta is a very efficient coolant, and some models show that the critical metallicity falls to $[{\rm Z/H}]_{\rm crit} \approx -6$ if such dust is produced efficiently.

Regardless of its precise value, the small critical metallicity is easy to achieve. We have seen that a single pair-instability supernova can enrich an entire halo as well as some portion of the IGM. Typical explosions generate $M_{{\rm SN},X} \sim 10\ M_\odot$ of C or $\sim 30\ M_\odot$ of O. A single supernova therefore enriches its host to a carbon abundance $\sim 3 \times 10^{-3} (M_{\rm SN,C}/10\ M_\odot)(10^6\ M_\odot/M)$ times the solar value (and a comparable level for oxygen). Thus, provided only that mixing is efficient, a single supernova suffices to shift star formation in its host halo – and possibly its close neighbors – into the Population II channel.

The above arguments show that fragmentation can occur in low metallicity environments, but they do not determine the actual spectrum of mass fragments. That is highly uncertain, but it is still likely to be skewed to significantly higher masses than today. The same arguments as in §5.2.4, in which the CMB sets the temperature floor for the cooling gas, will apply to these enriched clumps as well, setting the characteristic mass scale to be a few tens of solar masses – still well into the high-mass regime.

Nevertheless, the transition to Population II is a crucial milestone in the history

of the Universe. The arguments in this section suggest that, if mixing was efficient, it took place very soon after the first star in each virialized halo exploded.

6.5.1 Blastwaves in an Expanding Universe

A crucial point to understand about metal enrichment is that it *must* be highly inhomogeneous, because the metals are produced at discrete sites (star-forming halos) and must be advected with hydrodynamic flows, which typically move rather slowly by cosmological standards. Thus, the transition from Population III to Population II is likely to had large spatial fluctuations; in principle, Population III star formation could persist to late times, if the IGM enrichment timescales are very long and if new halos virialize and cool. In this section, we will consider how galactic winds (or other flows) can distribute this material around the Universe.

Although the simple models of §6.4.2 provide some intuition, they do not directly apply to cosmological blastwaves, which propagate into an expanding medium whose density decreases with time. However, it is easy in this case to estimate the maximum distance to which the shock can reach: as in a uniform, static medium, the wave will sweep the matter before it into a thin shell. But in the cosmological setting, the shell will continue to expand in comoving coordinates only while its velocity *relative to the Hubble flow* is positive – after that, the shell will simply be dragged along with the Hubble flow. The drag from swept up material will continue to decelerate it until its velocity matches the Hubble flow, which occurs at the asymptotic (proper) radius $R_{\rm E}$; the final kinetic energy is then $M_s[H(z_i)R_{\rm E}]^2/2$. (Here z_i is the initial time of the explosion.) Some of this energy comes from the explosion energy E, but in contrast to the static medium, the initial configuration also contains some kinetic energy from the Hubble flow. Integrating outward to $R_{\rm E}$, this initial energy is $3M_s[H(z_i)R_{\rm E}]^2/10$. If we assume that the expansion is rapid compared to the Hubble time, the maximum comoving size is therefore

$$R_{E,\rm com} \sim \left[\frac{E}{\rho_b(z_i)H^2(z_i)}\right]^{1/5}(1+z) \tag{6.38}$$

$$= K_{E,\cos}\left(\frac{GE}{H_0^4\Omega_b\Omega_m^2}\right)^{1/5}(1+z_i)^{-1/5}, \tag{6.39}$$

where $K_{E,\cos} \sim 10^{1/5}$ is a constant of order unity that depends on how much energy is transformed into thermal energy or kinetic energy. Note the similarity to the Sedov-Taylor-von Neumann scaling, with $t \sim 1/H(z)$.

In fact, for a perfectly adiabatic shock in a matter-dominated Universe with $\Omega_b \ll \Omega_m$, a self-similar blastwave that mirrors the Sedov-Taylor-von Neumann solution forms. In this limit, the constant $K_{\cos} = (32\pi/3)^{1/5}K_{\rm STV}$. The blastwave also expands at a rate $R_{\rm com} \propto \tau^{2/5}$, where at high redshifts,

$$\tau(z) \approx \frac{2}{\sqrt{\Omega_m}H_0}[(1+z_i)^{1/2} - (1+z)^{1/2}] \tag{6.40}$$

Once radiative cooling in the shell and eventually the bubble interior become important, the expansion slows down. We can estimate the final size of a bubble in which cooling is extremely efficient by repeating the above argument, but with

momentum conservation as our guiding principle rather than energy conservation. Writing the total impulse as E/c, we obtain

$$R_{p,\rm com} \sim K_{\rm cos,p} \left[\frac{E/c}{\rho_b(z_i)H(z_i)}\right]^{1/4} (1+z) \tag{6.41}$$

$$= K_{p,\rm cos} \left(\frac{GE/c}{H_0^3 \Omega_b \sqrt{\Omega_m}}\right)^{1/4} (1+z_i)^{-1/8}, \tag{6.42}$$

where $K_{p,\rm cos} \sim 8^{1/4}$.

To put these estimates in the context of star-forming halos, we use the notation of §6.4.1 and write the energy released by a halo of mass M as $E = f_\star \omega_{\rm SN} M_g$ and the momentum input as $E/c = (\epsilon f_\star M_g c^2)/c$, where $M_g = (\Omega_b/\Omega_m)M$. Then

$$R_{E,\rm com} \sim 1.2 \left(\frac{\omega_{\rm SN}}{10^{49}\ {\rm erg}\ M_\odot^{-1}} \frac{f_\star}{0.1} \frac{M}{10^8\ M_\odot}\right) (1+z)^{-1/5}\ {\rm Mpc}, \tag{6.43}$$

$$R_{p,\rm com} \sim 0.2 \left(\epsilon_3 \frac{f_\star}{0.1} \frac{M}{10^8\ M_\odot}\right) (1+z)^{-1/8}\ {\rm Mpc}. \tag{6.44}$$

It follows that the maximal comoving *volume* enriched by halos scales as $V_E \propto (f_\star M)^{3/5}$ or $V_p \propto (f_\star M)^{3/4}$. In either case, this is sublinear, showing that low-mass halos are much more efficient at enriching the IGM than massive ones.

In practice, these maximal radii can be substantial overestimates since they neglect the gravitational attraction of the host halo; see §6.4.1.

To follow the time evolution in detail one must track the energy reservoir driving the wind. Numerical calculations show that cosmological blastwaves develop shells even more rapidly than their counterparts in static media. The equation of motion for a shell is then

$$\ddot{R} = \frac{4\pi R^2}{M_s}(p - p_{\rm IGM}) - \frac{G}{R^2}[M(R) + M_s/2] + \Omega_\Lambda(z) H^2(z) R - \frac{\dot{M}_s}{M_s}(\dot{R} - HR), \tag{6.45}$$

where M_s is the shell mass, $\dot{M}_s = 4\pi R^2 \rho_b (\dot{R} - HR)$ is the rate at which mass is swept up, p is the pressure of the bubble interior, $p_{\rm IGM}$ is the ambient pressure of the IGM, and $M(R)$ is the mass enclosed within the wind (including both dark matter and any baryonic remnants). The first term is the pressure force from the hot interior, the second involves the gravitational deceleration due to the interior mass (and the shell itself), the third is the acceleration due to the cosmological constant (which can be ignored at high redshifts), and the drag force from swept-up material. This must be supplemented with an equation for the energy of the bubble interior,

$$\dot{p} = \frac{L}{2\pi R^3} - 5p\frac{\dot{R}}{R}. \tag{6.46}$$

Here the second term is the pdV work from expanding the shell, while the first represents energy inputs or losses. These include the energy source powering the wind and Compton cooling (which usually dominates at the low bubble densities once the winds propagate into the IGM).

There is one additional subtlety in the cosmological case: the shell treatment assumes that the ambient material is accelerated to the shell velocity through inelastic

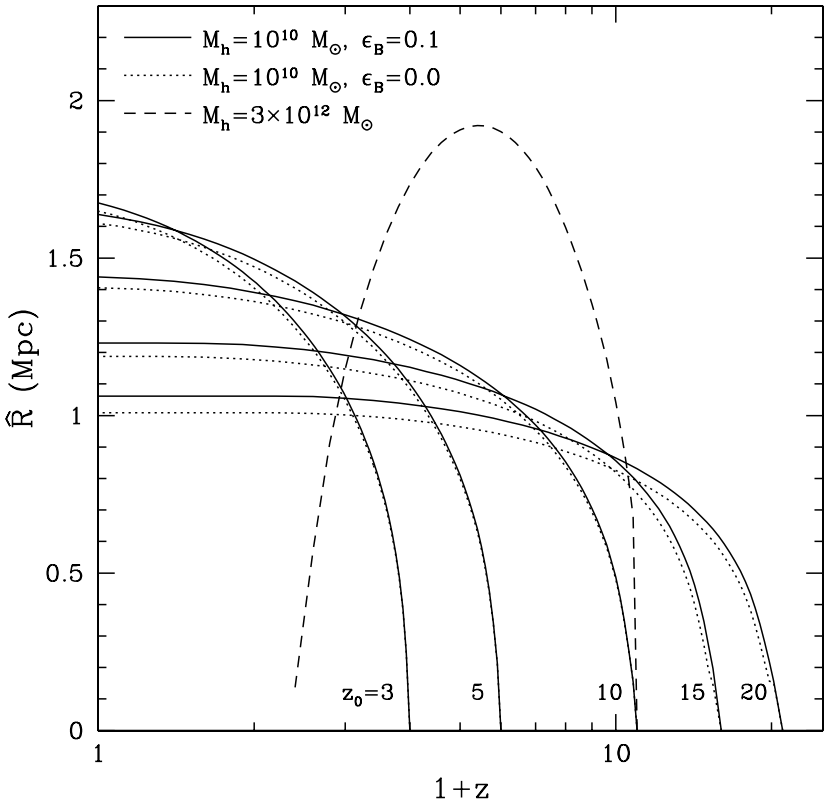

Figure 6.8 **Placeholder figure.** Shell sizes as a function of redshift for several different model halos.

collisions. L also must account for the energy dissipated in this process. If the shell cooling time is short, most of it will be lost (the static medium solutions described in §6.4.2 implicitly take this limit), but some may be transmitted to the bubble interior through turbulence if it is not lost in cooling. We let f_d be the fraction of this energy transmitted to the bubble interior; then L includes a term

$$L_d = f_d \dot{M}_s (\dot{R} - HR)^2/2. \tag{6.47}$$

Figure 6.8 shows solutions to these shell expansion equations for several model halos and winds. We show winds launching at $z = 30,\ 20,\ 15$, and 10, from halos with $T_{\rm vir} = 10^4$ K. In all cases we assume an instantaneous burst of star formation with $f_\star = 10^{-3},\ 10^{-2}$, and 0.1 from bottom to top within each set. We halt the expansion and assume that the bubbles are "frozen-in" to the Hubble flow once their expansion velocity falls below that limit. Note the fairly long times for the bubbles to reach these limiting sizes and the sublinear dependence of the limit on the total energy input.

Figure 6.9 compares the asymptotic bubble sizes as a function of input energy (in this case parameterized as halo mass, with $f_\star = 0.1$ held constant) in this detailed calculation to the maximal limit from equation (6.43), based on energy conservation, and the minimal limit from equation (6.44), based on the total impulse. The upper and lower solid lines adopt $f_d = 0$ and 1, respectively – these two curves therefore bracket the real bubble expansion, since the shell is expected to cool partway through the expansion. Note that the energy limit underestimates the final sizes because it does not account for the expanding Universe, which allows bubbles to grow farther both by the decelerating expansion and decreasing density. (On the other hand, Figure 6.8 shows that it takes a substantial amount of time to reach the maximal limit, so during the high-redshift era of interest the estimate is reasonably accurate.)

The numerical results turn over at high masses, because the gravitational potential well of the host traps the wind. Typically this occurs before the wind escapes far into the IGM, so there is a severe cutoff in the maximum size – recall that the gravitational binding energy scales as M^2, while the available energy only goes like M. This, together with the $V \propto E^{3/5} \propto (f_\star M)^{3/5}$ scaling of the enriched volume, mean that the smallest halos are likely the most important for chemical enrichment, unless the star formation efficiency itself decreases strongly at low halo masses.

6.5.2 Metals in the Intergalactic Medium

Given the fate of a wind bubble around any individual source, it is straightforward to estimate the fraction of space filled by these bubbles. Defining $V(m, z)$ to be the volume filled by a bubble blown by a halo of mass m at redshift z,[vi] we integrate over the halo mass function:

$$Q'_e(z) = \int_{M_{\rm min}}^{\infty} dM\, n(M, z) V(M, z), \tag{6.48}$$

where the integration extends over all star-forming halos. The resulting Q'_e is the total volume filled by all the bubbles, not accounting for overlap. If the bubbles were randomly distributed, and if overlapping winds did not aid each other's expansion, the true filling fraction of wind material would be $Q_e = (1 - e^{-Q'_e})$.

This simple estimate has an important shortcoming: it ignores the clustering of these galaxies. In reality, high-redshift galaxies form close to each other along intersections of sheets and filaments in the cosmic web. Their wind bubbles therefore tend to overlap rather than fill new space. Because $V \propto E^{3/5}$, multiple sources contributing to a single bubble are *less* efficient than individual sources generating their own bubbles, so clustering will tend to decrease the filling fraction of the enriched material.

Figure 6.10 shows some example enrichment histories with our simple approach. We consider two different models: a maximal estimate in which star formation in

[vi] In reality, there will be quite a bit of scatter in this relationship, as halos (even of the same mass) form and grow with different merger and star formation histories. For the simple estimates here we ignore this scatter, though it can be easily followed with numerical simulations.

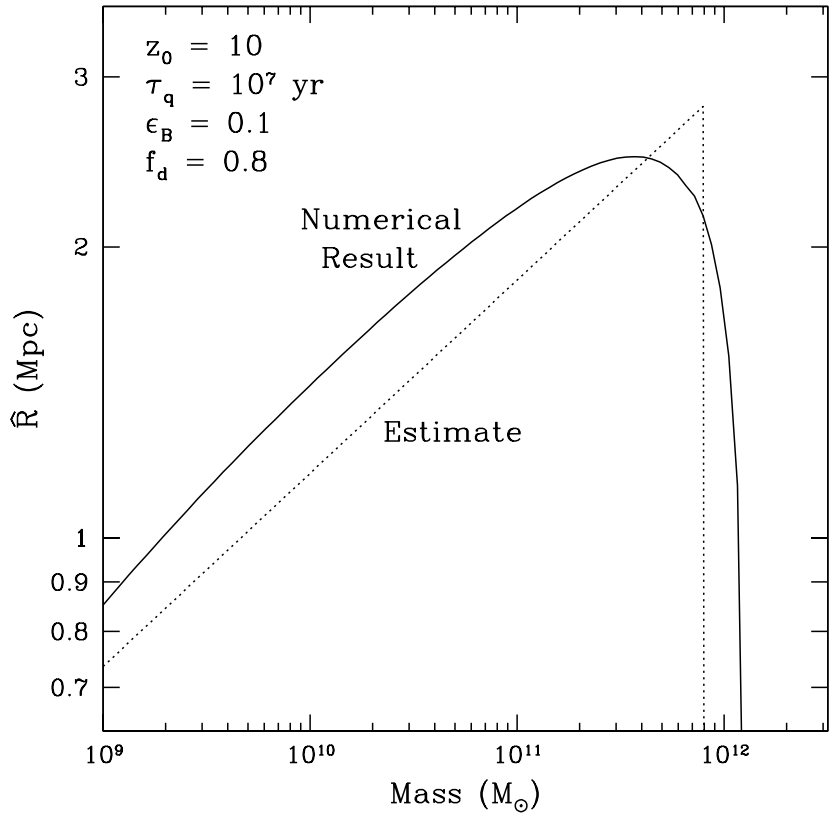

Figure 6.9 **Placeholder figure.** Comparison of maximal wind sizes in the detailed model of eq. (6.45) (solid lines) with the estimates of eq. (6.43) and (6.44) (dashed and dotted lines, respectively) based on energy and momentum conservation. The upper and lower solid lines show solutions with $f_d = 1$ and 0, respectively. In reality, the shell likely cools partway through the expansion, so full integrations lie somewhere between these curves. The numerical results turn over at high masses because the wind cannot escape the gravitational potential well of the host halo.

small halos through H_2 cooling is efficient (with all halos with $T_{\rm vir} > 1000\,{\rm K}$ having $f_\star = 0.1$), a more conservative model in which star formation in such halos is very inefficient (with $f_\star = 10^{-3}$ unless $T_{\rm vir} > 10^4$ K). In both cases we show the numerical integrations (with the solid lines, setting $f_d = 1$) as well as the energy and momentum estimates (with the dotted and dashed lines, respectively).

Figure 6.10 shows that, in order for a large fraction of space to be filled with heavy elements by $z \sim 6$, much of those metals must come from the shallow potential wells of very small halos, which must produce stars very efficiently. Indeed, if supernova and photoionization feedback is as efficient as our earlier estimates suggest, it seems implausible to expect such halos to be able to convert 10% of their baryons into stars. Thus, metal enrichment in these early phases seems likely to be very patchy, with important consequences for structure formation (see §6.6).

In galaxies that were likely responsible for most of the metal enrichment, both supernova winds and radiation pressure from hot stars contributed to powering the outflows. The former ultimately provide more energy for the outflow, but much of that energy may be lost as the supernova blast waves propagate through the dense ISM of the galaxies. The momentum inputs from the two channels are comparable for typical IMFs, so even if supernova remnant cooling is efficient winds from starbursts should be able to enrich a few percent of the IGM.

Unfortunately, numerical cosmological simulations currently lack the dynamic range to model the launch of these winds and their propagation through the IGM (because the shells cannot be resolved), although simulations of individual galaxies are beginning to examine outflow dynamics in detail. In large-scale structure simulations, winds are launched by hand with a parameterized model; they are then tracked as they propagate through the IGM in the momentum-dominated limit. Such numerical simulations also show that plausible models for winds from halos above the 10^4 K cooling threshold can enrich only a few percent of the IGM. As the winds continue to expand at later times, this fraction increases, but many models predict that much of the IGM remains pristine even to late times.

The mean metallicity of these enriched regions follows easily from the above models with only one additional parameter: the fraction $f_{\rm met}$ of metals that are ejected in the wind. This is usually parameterized with the *mass-loading factor* η which describes how much material escapes the galaxy in units of the star formation rate, $\eta = \dot{M}_w/\dot{M}_\star$. According to the model described in §6.4.1, $\eta \sim \sigma_0/\sigma$ (see Eq. 6.32), assuming that the momentum input rate simply scales with the star formation rate and that the final velocity of the winds is just a few times the escape velocity of the halo. Observations of low-redshift starbursts are consistent with this simple relation if $\sigma_0 \sim 300{\rm km\,s^{-1}}$, though we note that the proportionality constant depends on the IMF and may get *larger* if the IMF is top-heavy at high redshifts. On the other hand, this provides yet another reason why small galaxies more efficiently enrich the IGM with metals, as $\eta \propto \sigma^{-1} \propto M^{-1/3}$.

If we then assume that the metals are perfectly mixed inside the galaxy, this implies that a fraction $\eta f_\star$ of the metals produced in each galaxy are ejected into the IGM. This material is then diluted by a factor $\sim Q_e f_{\rm coll}$; thus, the mean IGM

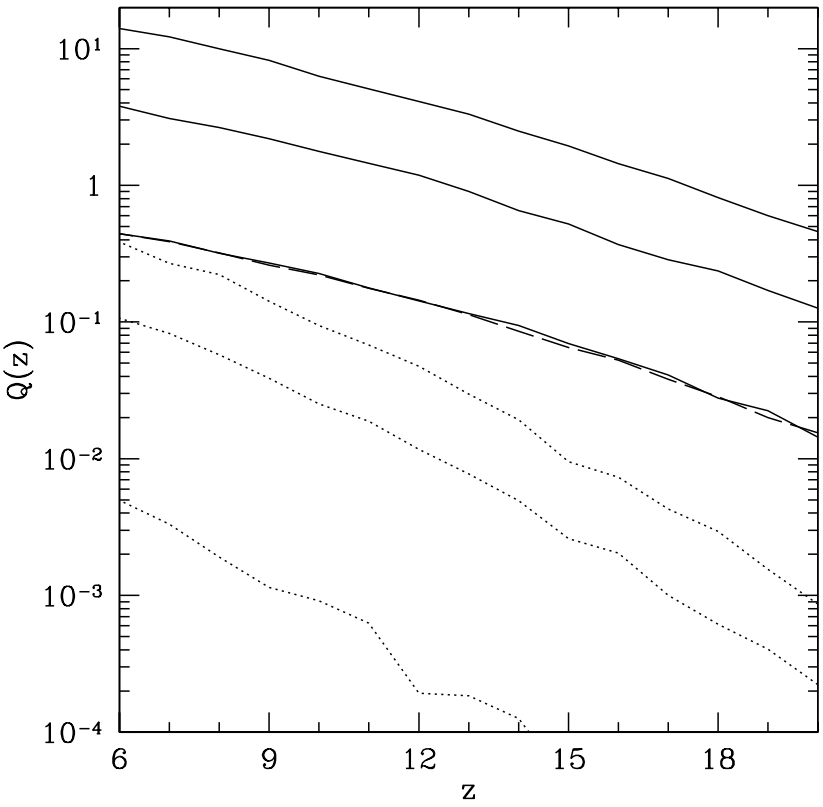

Figure 6.10 **Placeholder figure.** Filling factor of wind-enriched regions in different models of star formation and wind expansion. The upper and lower set of curves show models with efficient star formation in halos below the atomic cooling threshold (setting $f_\star = 0.1$ in all halos with $T_{\rm vir} > 1000$ K) and in which it is not (setting $f_\star = 10^{-3}$ in halos with $T_{\rm vir} < 10^4$ K). The solid curves show numerical calculations using eq. (6.45), while the dotted and dashed curves use estimates from energy and momentum conservation, respectively (Eqs. 6.43 and 6.44).

metallicity will be

$$Z_{\rm IGM} \sim 10^{-3} \langle \eta \rangle \left(\frac{f_\star}{0.1} \frac{f_{\rm coll}}{0.01} \right) Z_{\rm gal}, \tag{6.49}$$

where $Z_{\rm gal}$ is the mean metallicity of material inside of galaxies and $\langle \eta \rangle$ is averaged over the entire galaxy population. The mean metallicity of *enriched* regions will be larger by $\sim Q_e^{-1}$. Because Q_e should increase with $f_{\rm coll}$, this shows that the metallicity of enriched regions is likely to be above the critical threshold for the transition to Population II star formation in most plausible scenarios.

An alternative empirical estimate of the IGM metallicity follows by observing the total density of stars (which, assuming an IMF, translates into a total metal yield). Type II supernovae from high-mass stars forming in a typical Salpeter IMF process $\approx 2.4\%$ of the stellar mass into metals. Using the observed stellar mass estimates at $z \sim 2$, this implies that the IGM should have $Z \sim (1/30)\ Z_\odot$ at that redshift.

For a similar constraint at higher redshifts, we can calibrate the stellar mass to the number of ionizing photons produced per baryon, which will let us gauge the overall level of enrichment near the time of reionization. We let Q be the number of ionizing photons reaching the IGM per hydrogen atom, $Q \approx N_\gamma f_{\rm esc} f_\star f_{\rm coll}$, where N_γ is the number of ionizing photons produced per baryon in stars (~ 4000 for a Salpeter IMF) and $f_{\rm esc}$ is the fraction of these photons that escape their host galaxy into the IGM; we will discuss these parameters in more detail in §8. Meanwhile, the mean metallicity implied by these stars is $Z/Z_\odot \sim 1.3 f_\star f_{\rm coll}$, where the factor $f_\star$ is the conversion from the 2.4% metal yield to solar metallicity (with 1.89% of the mass in metals). Thus,

$$Z \sim 3 \times 10^{-3} Q \left(\frac{400}{N_\gamma f_{\rm esc}} \right)\ Z_\odot. \tag{6.50}$$

Again, the mean metallicty in enriched regions will be a factor Q_e^{-1} larger.

Our primary tools for constraining these winds are metal-line systems in the Lyman-α forest (see §4.5). Metals seem ubiquitous in the high-column density systems that may be associated with virialized objects, which implies that such halos are highly enriched. This is not surprising, since the first stars in any halo are themselves likely enrich the hosts' material to substantial levels. More interesting is the wide scatter in the metallicity of lower-density regions. The estimate in equation (6.49) is reasonably close to the observed metallicities of these systems ($Z \sim 10^{-3} Z_\odot$), so careful studies of IGM metal lines over time may shed light on winds and other outflows. In particular, even at $z > 6$, these enriched regions will produce measurable absorption in quasar or GRB spectra, although identifying each line's origin may be difficult (see §4.6).

However, only $\sim 10\%$ of these metals predicted by measuring the stellar mass of the Universe have actually been observed: the remainder may be buried inside additional galaxy populations or in the IGM. If the latter, this suggests that the enrichment may indeed be very widespread, at least by $z \sim 2$–3.

As we will see in §9.5, these winds likely also play a crucial role in regulating star formation within galaxies, and their parameters can therefore be estimated

not only through IGM metallicity measurements but also by comparison to galaxy luminosity functions, metallicities, and other properties. This provides another observational handle on winds and, indirectly, chemical enrichment processes.

6.6 THE FIRST GALAXIES

In Chapter 5, we discussed the physics of primordial star formation. Although there are many unanswered questions, the problem of first *star* formation is a tractable one: the initial conditions are well-posed and the physics (dark matter and baryonic collapse, chemistry of the primordial gas, accretion disk formation, and radiative feedback) is straightforward enough that one can at least imagine solving the problem in full.

However, in this chapter we have examined the myriad feedback mechanisms generated by these stars and their descendants. As soon as the first stars form, these processes complicate matters immensely, and it is extremely difficult to imagine building a picture of the subsequent generations of star formation from first principles – there are simply too many uncertain parameters driving each one. Nevertheless, the underlying physics of each process is relatively straightforward, and from detailed studies of each individual process we can build some intuition for how the interplay may proceed.

Such "global" formulations are coming into focus for the transformation of the first stars to the first *galaxies*. We define a "galaxy" as a gravitationally-bound system of stars embedded in a dark matter halo and exhibiting *sustained* star formation (even if at a low level) over cosmological time periods (i.e., a substantial fraction of the Hubble time). This definition requires: (i) a virialized dark matter halo able to accrete baryons (hence $M > M_{\rm fil}$); (ii) efficient cooling in the baryons (above a critical virial temperature $T_{\rm min}$ that depends on the chemistry of the constituent gas); (iii) sufficient mass to be stable against feedback from its own stars; and (iv) sufficient mass to be stable against feedback from neighboring halos.

Here we will describe a plausible scenario for how such objects can appear at high redshifts. It should be obvious, however, that though this represents a "best guess" given present theoretical investigations, the lack of observational constraints on this expectation likely means that it is at best partially correct. Nevertheless, it provides a coherent synthesis of the concepts we have discussed and is a useful baseline paradigm for future work. Figure 6.11 illustrates the following evolutionary stages graphically and identifies some of their key points:

1. The first stars form inside halos cooled by molecular hydrogen, with characteristic masses determined by the chemistry of H_2 cooling (see §5.1.2 and Fig. **??**), with $T_{\rm vir} > 1000\,{\rm K}$. Massive Population III.1 stars form at the center of these halos after cooling to low temperatures. The key question is whether the gas cloud fragments before the material accretes onto the protostar. If not, the final mass is likely regulated by radiative feedback (with $M_\star > 100\,M_\odot$); alternatively, the first protostar's accretion disk is the most likely site for fragmentation, and the characteristic mass may be several times

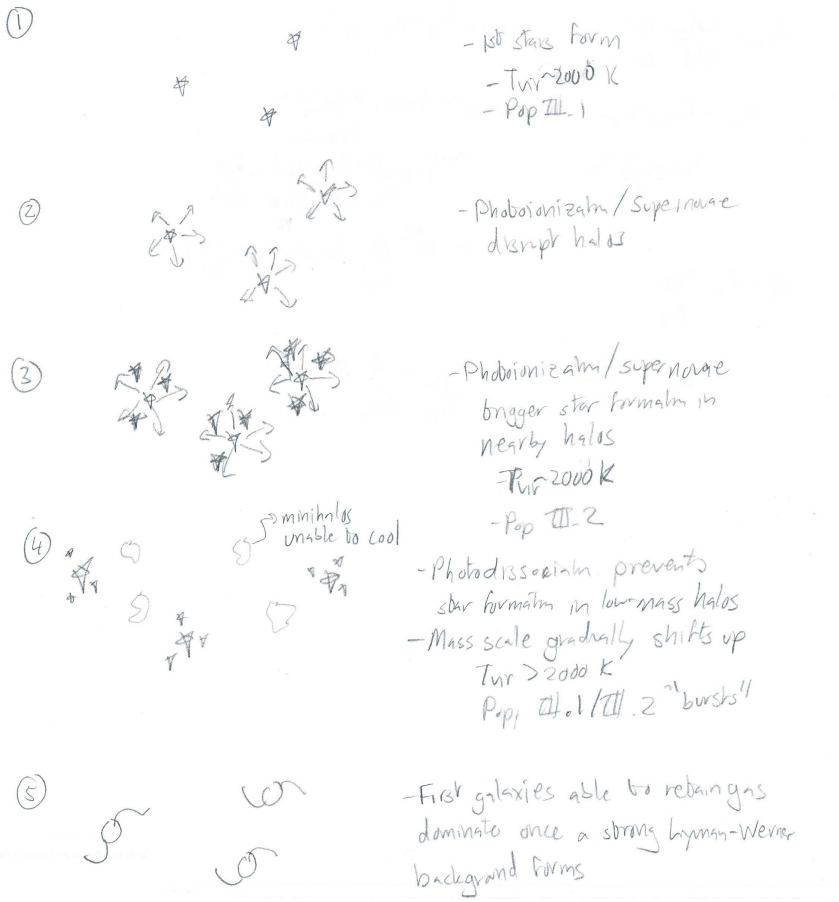

Figure 6.11 Stages in a plausible scenario for the birth of the first stars and galaxies (see text for details). *(1)* The first Population III.1 stars form in small halos via H_2 cooling. *(2)* These stars empty their hosts of gas via photoevaporation and supernova blastwaves. *(3)* This feedback triggers Population III.2 star formation in nearby minihalos. *(4)* The Lyman-Werner background from these stars suppresses star formation in small minihalos, gradually increasing the characteristic mass scale of star-forming objects. *(5)* The first self-sustaining galaxies eventually form in halos above the atomic cooling threshold, $T_{\rm vir} \sim 10^4$ K.

smaller.

2. The first star (or star cluster) exerts extremely strong feedback on its host halo's gas. The H II region created by a very massive Population III.1 star evaporates any diffuse gas in the central regions of the halo (§6.3.1), and the star's death as a supernova will trigger a blastwave that quickly clears out the rest of the gas (provided, of course, that the star does not collapse directly to a black hole without generating an explosion; §6.4.3), and it enriches the entire halo with heavy elements. Dense clumps well on their way to star formation may survive this feedback (and in fact the shock compression may even speed up their collapse), but nothing else will. The feedback will be less severe, but still substantial, if Population III.1 stars are less massive. Nevertheless, *Population III.1 star formation in any individual halo may only occur in a single rapid burst.*

3. These same feedback mechanisms also operate on somewhat larger scales, as the H II region and supernova blastwave are able to penetrate to $\sim$ kpc scales. Any nearby halos will therefore be subject to the same effects: those that have not yet collapsed to high densities will have their baryons evaporated at high entropies, while those already dense enough to self-shield from the ionizing radiation will likely have their star formation accelerated (§**??**). However, because these systems will form their stars from ionized gas, the enhanced HD chmistry will lead to more efficient cooling and hence (probably) a smaller characteristic mass of Population III.2 stars (see §5.3). (Note that, because supernova blastwaves travel much slower than H II regions, it is very possible for this triggered star formation to be metal-free.) Still, even with this positive feedback, the Population III.1 and Population III.2 stars in a given cluster of minihalos will form temporally close together (as otherwise the clumps would have been photoevaporated), leading to "bursts" of Population III stars followed by long pauses as the halos re-accrete their gas. Figure 6.12 illustrates some of the complexity of this stage: note the several nearby stars that form and the complicated morphology of the molecular gas catalyzed by the presence of the H II regions.

4. Feedback also operates on larger scales. All Population III stars produce photons in the Lyman-Werner that photodissociate H_2. As more stars form, the Lyman-Werner background increases, gradually raising the critical virial temperature for cold gas formation inside minihalos (§6.1.5 and Fig. 6.3). Because more massive halos are also more rare, this will tend to self-regulate the global rate of star formation.

5. Eventually, the Lyman-Werner background will become intense enough to choke off Population III star formation in pristine minihalos entirely. Then star formation will shift to halos with $T_{\rm vir} > 10^4$ K, where H I is ionized by the virial shock and atomic cooling is efficient (see Fig. 5.1). Most likely these halos will have had progenitors that formed Population III stars, in which case they will already be pre-enriched with metals and begin to

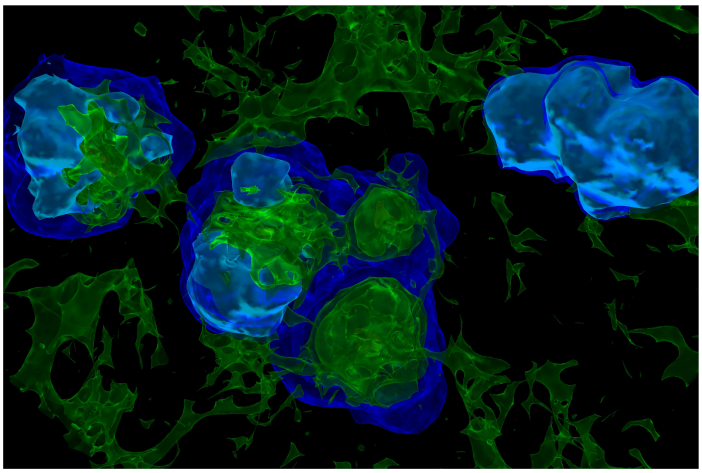

Figure 6.12 Results from a numerical simulation of the formation of a metal-free stars and their feedback on the surrounding environment. Radiative feedback around the first star involves ionized bubbles (light grey) and regions of high molecule abundance (medium grey). The large residual free electron fraction inside the relic ionized regions, left behind after the central star has died, rapidly catalyzes the reformation of molecules and a new generation of lower-mass stars. Figure credit: Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F. *Nature* **459**, 49 (2009).

form Population II stars. It is possible, however, that some such halos will form without stars inside their progenitors (perhaps because they form relatively late and so have star formation suppressed by the Lyman-Werner background). In that case, their initially-ionized gas will cause Population III.2 star formation.

6. Systems with $T_{\rm vir} > 10^4$ K can also maintain reasonable (though still small) star formation rates without completely disrupting their gas supplies (see Eq. 6.25). *It is therefore this "second-generation" of star-forming halos that host the first sustained galaxies.*

7. Nevertheless, feedback continues to be important in regulating galaxy formation at later times. Winds and outflows are likely crucial for regulating star formation inside galaxies (see §9.5), and photoheating from ionizing photons in the IGM will gradually increase the Jeans mass and so increase the minimum mass scale for galaxy formation (see §8.10, where we will discuss this topic in detail.)

The transition to star formation in long-lived galaxies likely occurred long before the Universe was reionized. The intensity of the Lyman-Werner background can be estimated as

$$J_{\rm LW} \sim \frac{cn_{\rm LW}}{4\pi}\left(\frac{h\nu}{\Delta\nu_{\rm LW}}\right) \tag{6.51}$$

where $n_{\rm LW}$ is the number density of photons in the Lyman-Werner band and $\Delta\nu_{\rm LW}$ is the band's width in frequency space. We then write $n_{\rm LW} \sim f_{\rm LW/ion}Q/f_{\rm esc}n_H$, where $f_{\rm LW/ion}$ is the number of Lyman-Werner photons produced per ionizing photon by stars (which is ~ 0.1 for very massive Population III stars or near unity for Population II stars), Q is the number of ionizing photons that escape into the IGM, and $f_{\rm esc}$ is the fraction of all ionizing photons that manage to escape in this way. Then

$$J_{\rm LW,21} \sim 100Q\left(\frac{0.1}{f_{\rm esc}}\frac{f_{\rm LW/ion}}{0.1}\right)\left(\frac{1+z}{20}\right)^3. \tag{6.52}$$

Lyman-Werner photons suppress H_2 cooling completely when $J_{\rm LW,21} > 1$, which should occur long before enough ionizing photons are produced to reionize the IGM. Thus, it seems very likely that the primary sources responsible for reionizing the Universe were long-lived galaxies, rather than the bursty minihalos in which the first stars themselves formed.

Although this is a very plausible picture consistent with detailed theoretical work, there are a number of points at which seemingly minor choices may dramatically alter the results. We list several here to give a flavor for the uncertainties:

- If fragmentation is efficient in accretion disks composed of primordial stars, the first halos would form clusters of moderately sized stars rather than single very massive stars. The resulting feedback would be less efficient, potentially allowing gas to remain in halos somewhat below the usual $T_{\rm vir} \sim 10^4$ K atomic cooling threshold. The mass scale of the first galaxies would shift downward.

- If Population III stars form in the mass ranges 40–100 $M_\odot$ or 140–260 $M_\odot$, they will die by exploding rapidly to black holes without explosions. This would allow their halos to retain more of their gas, with only the photoevaporation feedback to contend with, and allow sustained star formation to continue in low-mass halos. Moreover, they would not enrich their environments (except perhaps weakly through stellar winds), allowing Population III.1 and III.2 to persist for longer timescales – possibly even to the atomic cooling threshold.

- If the shells that form at the edges of supernova blastwaves are gravitationally unstable, they can fragment and form stars as well. If the fragmentation scale is small, these could even be long-lived stars who exert relatively small feedback.

- If black holes form abundantly and accrete gas efficiently from binary star companions or the ISM (see § 7), then their X-ray background increases the free electron fraction inside halos, promoting H_2 formation and possibly counteracting photodissociation from Lyman-Werner photons. This would allow much more rapid primordial star formation in low-mass halos.

- The consequences of enrichment inside minihalos has been largely unexplored, because the gas is expected to be expelled. But if some is retained, the metals allow rapid cooling and hence more efficient star formation than H_2. This too could lead to smaller galaxies.

Obviously there is a great deal of uncertainty in how the first stars will grow into the first galaxies – most likely, observations will be necessary to settle the question. However, in closing we stress that most of the underlying physics is well-understood in isolation and has many applications to other areas of astrophysics. It is the complex interplay of the processes we have described here that makes the problem challenging and exciting to explore observationally.

Chapter Seven

Supermassive Black holes

Why did the collapsed matter in the Universe end up making galaxies and not black holes? One would have naively expected a spherical collapse to end with the formation of a point mass at its center. But, as it turns out, tides from neighboring objects torque the infalling material and induce non-sphericity and some spin into the final collapse. The induced angular momentum prevents the gas from reaching the center on a direct plunging orbit. After the gas cools and loses its pressure support against gravity, it instead assembles into a disk in which the centrifugal force balances gravity. The finite size of the luminous region of galaxies is then dictated by the characteristic spin acquired by galaxy halos, which typically corresponds to a rotational velocity that is $\sim 5\%$ of the virial circular velocity, with a negligible dependence on halo mass. This does not imply that no gas accumulates at the center. In fact, galactic spheroids are observed to generically harbor a central black hole, whose formation is most likely linked to a small mass fraction the galactic gas ($< 0.1\%$) which has an unusually low amount of angular momentum. The small mass fraction of the central black holes implies that their gravitational effect is restricted to the innermost cusp of their host galaxy. Nevertheless, these central black holes are known to have a strong influence on the evolution of their host galaxies. This state of affairs can be easily understood from the fact that the binding energy per unit mass in a typical galaxy correspond to velocities v of hundreds of $\mathrm{km\ s^{-1}}$ or a fraction $\sim (v/c)^2 \sim 10^{-6}$ of the binding energy per unit mass near a black hole. Hence a small amount of gas that releases its binding energy near a black hole can have a large effect on the rest of the gas in the galaxy.

The growth of supermassive black holes is intimately linked to the hierarchical growth of their host galaxies. Figure 7.1 shows the evolution of the luminosity function of quasars at different observed wavelengths in the redshift interval $z = 2$–5. The inferred growth in the comoving mass function of black holes along with its integral over all black hole masses (i.e. the comoving mass density) are shown in Figure 7.2. The highest redshift quasar known is ULAS J1120+0641 at $z = 7.085$ (only 0.77 Gyr after the Big Bang), with a bolometric luminosity of $6.3 \times 10^{13} L_\odot$ and an estimated black hole mass of $2 \times 10^9 M_\odot$.

We start this chapter with a short introduction to the properties of black holes in general relativity.

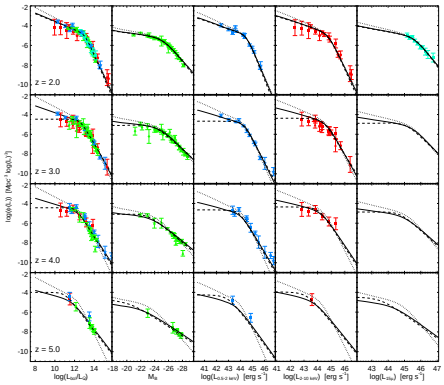

Figure 7.1 Redshift evolution of the luminosity function of quasars at different observed wavelengths, B-band (center-left panels), soft X-rays ($0.5-2$ keV) (center), hard X-rays ($2-10$ keV) (center-right; red), and mid-IR ($15\,\mu$m) (right; cyan). The left panels show the distribution of bolometric luminosities (integrated over all wavelength). Lines show the best-fit evolving double power-law model to data points at all redshifts (solid), the best-fit model at the given redshift (dashed), and the best-fit model that allows only the break luminosity to evolve (dotted). Figure credit: P. F. Hopkins, G. T. Richards, & L. Hernquist, *Astrophys. J.* **654**, 731 (2007).

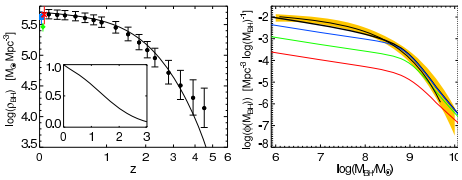

Figure 7.2 *Left panel:* the black hole mass density of quasars from the data (circles) and the best fit luminosity function (solid line). The inset shows the fraction of the mass density at $z = 0$ ($\rho_{\rm BH}(z)/\rho_{\rm BH}(0)$), on a linear scale. *Right:* the black hole mass function at $z = 0$ (thick black line) and $z = 1,\ 2,\ 3$ (from top to bottom). The shaded region shows the 1σ observational uncertainty. Figure credit: P. F. Hopkins, G. T. Richards, & L. Hernquist, *Astrophys. J.* **654**, 731 (2007).

7.1 BASIC PRINCIPLES OF ASTROPHYSICAL BLACK HOLES

In Newtonian gravity, the gravitational field at any radius outside a spherical mass distribution depends only on the mass interior to that radius. This results is also true in Einstein's General Relativity, where Birkhoff's theorem states that the only vacuum, spherically symmetric gravitational field is that described by the static *Schwarzschild metric*,

$$ds^2 = -\left(1 - \frac{r_{\rm Sch}}{r}\right) c^2 dt^2 + \left(1 - \frac{r_{\rm Sch}}{r}\right)^{-1} dr^2 + r^2 d\Omega, \qquad (7.1)$$

where $d\Omega = (d\theta^2 + \sin^2\theta d\phi^2)$. The *Schwarzschild radius* is related to the mass M of the central (non-spinning) black hole,

$$r_{\rm Sch} = \frac{2GM}{c^2} = 2.95 \times 10^5 \ {\rm cm} \left(\frac{M}{1M_\odot}\right). \qquad (7.2)$$

The black hole horizon, $r_{\rm Hor}$ ($= r_{\rm Sch}$ here), is a spherical boundary from where no particle can escape. (The coordinate singularity of the Schwarzschild metric at $r = r_{\rm Sch}$ can be removed through a transformation to the *Kruskal* coordinate system $(r,t) \rightarrow (u,v)$, where $u = (r/r_{\rm Sch} - 1)^{1/2} e^{r/2r_{\rm Sch}} \cosh(ct/2r_{\rm Sch})$; $v = u \tanh(ct/2r_{\rm Sch})$.) The existence of a region in space into which particles may fall but never come out breakd time reversal symmetry that characterizes the equations of quantum mechanics. Any grander theory that would unify quantum mechanics and gravity must remedy this conceptual inconsistency.

In addition to its mass M, a black hole can only be characterized by its spin J and electric charge Q (similarly to an elementary particle). In astrophysical circumstances, any initial charge of the black hole would be quickly neutralized through the polarization of the background plasma and the preferential infall of electrons or protons. The residual electric charge would exert an electric force on an electron that is comparable to the gravitational force on a proton, $eQ \sim GMm_p$, implying $(Q^2/GM^2) \sim Gm_p^2/e^2 \sim 10^{-36}$ and a negligible contribution of the charge to the metric. A spin, however, may modify the metric considerably.

The general solution of Einstein's equations for a spinning black hole was derived by Kerr in 1963, and can be written most conveniently in the Boyer-Lindquist coordinates,

$$\begin{aligned} ds^2 = &-\left(1 - \frac{r_{\rm Sch} r}{\Sigma_k}\right) c^2 dt^2 - \frac{2jr_{\rm Sch} r \sin^2\theta}{\Sigma_k} c dt d\phi + \frac{\Sigma_k}{\Delta} dr^2 \\ &+ \Sigma_k d\theta^2 + \left(r^2 + j^2 + \frac{r_{\rm Sch} j^2 r \sin^2\theta}{\Sigma_k}\right) \sin^2\theta d\phi^2. \end{aligned} \qquad (7.3)$$

where the black hole is rotating in the ϕ direction, $j = [J/Mc]$ is the normalized angular momentum per unit mass (in units of cm), $\Delta = r^2 - r r_{\rm Sch} + j^2$, and $\Sigma_k = r^2 + j^2 \cos^2\theta$. The dimensionless ratio $a = j/(GM/c^2)$ is bounded by unity, and $a = 1$ corresponds to a maximally rotating black hole. The horizon radius $r_{\rm Hor}$ is now located at the larger root of the equation $\Delta = 0$, namely $r_+ = \frac{1}{2} r_{\rm Sch}[1 + (1 - a^2)^{1/2}]$. The Kerr metric converges to the Schwarzschild metric for $a = 0$. There is no Birkhoff's theorem for a rotating black hole.

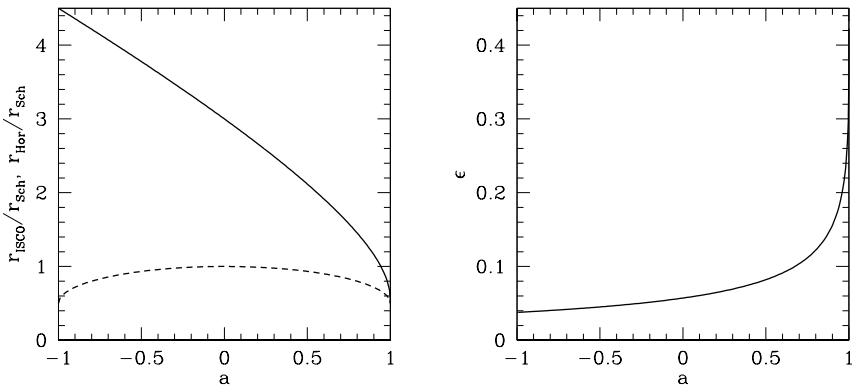

Figure 7.3 The left panel shows the radius of the black hole horizon $r_{\rm Hor}$ (dashed line) and the *Innermost Circular Stable Orbit (ISCO)* around it $r_{\rm ISCO}$ (solid line), in units of the Schwarzschild radius $r_{\rm Sch}$ (see equation 7.2), as functions of the black hole spin parameter a. The limiting value of $a = 1$ ($a = -1$) corresponds to a corotating (counter-rotating) orbit around a maximally-spinning black hole. The binding energy of a test particle at the ISCO determines the radiative efficiency ϵ of a thin accretion disk around the black hole, shown on the right panel.

Test particle orbits around black holes can be simply described in terms of an effective potential. For photons around a Schwarzschild black hole, the potential is simply $V_{\rm ph} = (1 - r_{\rm Sch}/r)/r^2$. This leads to circular photon orbits at a radius $r_{\rm ph} = \frac{3}{2} r_{\rm Sch}$. For a spinning black hole,

$$r_{\rm ph} = r_{\rm Sch} \left[1 + \cos\left(\frac{2}{3} \cos^{-1}[\pm a] \right) \right], \tag{7.4}$$

where the upper sign refers to orbits that rotate in the opposite direction to the black hole (retrograde orbits) and the lower sign to corotating (prograde) orbits. For a maximally-rotating black hole ($|a| = 1$), the photon orbit radius is $r_{\rm ph} = \frac{1}{2} r_{\rm Sch}$ for a prograde orbit and $2 r_{\rm Sch}$ for a retrograde orbit.

Circular orbits of massive particles exist when the first derivative of their effective potential (including angular momentum) with respect to radius vanishes, and these orbits are stable if the second derivative of the potential is positive. The radius of the *Innermost Circular Stable Orbit (ISCO)* defines the inner edge of any disk of particles in circular motion (such as fluid elements in an accretion disk). At smaller radii, gravitationally bound particles plunge into the black hole on a dynamical time. This radius of the ISCO is given by,

$$r_{\rm ISCO} = \frac{1}{2} r_{\rm Sch} \left\{ 3 + Z_2 \pm [(3 - Z_1)(3 + Z_1 + 2 Z_2)]^{1/2} \right\}, \tag{7.5}$$

where $Z_1 = 1 + (1 - a^2)^{1/3}[(1 + a)^{1/3} + (1 - a)^{1/3}]$ and $Z_2 = (3a^2 + Z_1^2)^{1/2}$. Figure 7.3 shows the radius of the ISCO as a function of spin. The binding energy

of particles at the ISCO define their maximum radiative efficiency because they spend a short time on their plunging orbit interior to the ISCO. This efficiency is given by

$$\epsilon = 1 - \frac{r^2 - r_{\rm Sch} r \mp j\sqrt{\frac{1}{2} r_{\rm Sch} r}}{r(r^2 - \frac{3}{2} r_{\rm Sch} r \mp 2j\sqrt{\frac{1}{2} r_{\rm Sch} r})^{1/2}}. \quad (7.6)$$

The efficiency changes between a value of $\epsilon = (1 - \sqrt{8/9}) = 5.72\%$ for $a = 0$, to $(1 - \sqrt{1/3}) = 42.3\%$ for a prograde (corotating) orbit with $a = 1$ and $(1 - \sqrt{25/27}) = 3.77\%$ for a retrograde orbit.

7.2 ACCRETION OF GAS ONTO BLACK HOLES

7.2.1 Bondi Accretion

Consider a black hole embedded in a hydrogen plasma of uniform density $\rho_0 = m_p n_0$ and temperature T_0. The thermal protons in the gas are moving around at roughly the sound speed $c_s \sim \sqrt{k_B T/m_p}$. The black hole gravity could drive accretion of gas particles that are gravitationally bound to it, namely interior to the radius of influence, $r_{inf} \sim GM/c_s^2$. The steady mass flux of particles entering this radius is $\rho_0 c_s$. Multiplying this flux by the surface area associated with the radius of influence gives the supply rate of fresh gas,

$$\dot{M} \approx \pi r_{\rm inf}^2 \rho_0 c_s = 15 \left(\frac{M}{10^8 M_\odot}\right)^2 \left(\frac{n_0}{1\ {\rm cm}^{-3}}\right) \left(\frac{T_0}{10^4\ {\rm K}}\right)^{-3/2} M_\odot\ {\rm yr}^{-1}. \quad (7.7)$$

In a steady state this supply rate equals the mass accretion rate into the black hole.

The explicit steady state solution to the conservations equations of the gas (mass, momentum, and energy) was first derived by Bondi (1952). The exact solution introduces a correction factor of order unity to equation (7.7). The solution is self-similar. Well inside the sonic radius the velocity is close to free-fall $u \sim (2GM/r)^{1/2}$ and the gas density is $\rho \sim \rho_0 (r/r_{\rm inf})^{-3/2}$. The radiative efficiency is small, because either the gas is tenuous so that its cooling time is longer than its accretion (free-fall) time or the gas is dense and the diffusion time of the radiation outwards is much longer than the free-fall time. If the inflowing gas contains near-equipartition magnetic fields, then cooling through synchrotron emission typically dominates over free-free emission.

A black hole that is moving with a velocity V relative to a uniform medium accretes at a lower rate than a stationary black hole. At high velocities, the radius of influence of the black hole would be now $\sim GM/V^2$, suggesting that the sound speed c_s be crudely replaced with $\sim (c_s^2 + V^2)^{1/2}$ in equation (7.7). A similar suppresion factor applies for the accretion of baryons onto dark matter halos, when the baryons have a net bulk velocity relative to the dark matter (see §???).

7.2.2 Thin Disk Accretion

If the inflow is endowed with rotation, the gas would reach a centrifugal barrier from where it could only accrete farther inwards after its angular momentum has been transported away. This limitation follows from the steeper radial scaling of the centrifugal acceleration ($\propto r^{-3}$) compared to the gravitational acceleration ($\propto r^{-2}$). Near the centrifugal barrier, where the gas is held against gravity by rotation, an accretion disk would form around the black hole, centered on the plane perpendicular to the rotation axis. The accretion time would then be dictated by the rate at which angular momentum is transported through viscous stress, and could be significantly longer than the free-fall time for a non-rotating flow (such as described by the Bondi accretion model). As the gas settles to a disk, the dissipation of its kinetic energy into heat would make the disk thick and hot, with a proton temperature close to the gravitational potential energy per proton $\sim 10^{12}\,{\rm K}(r/r_{\rm Sch})^{-1}$. However, if the cooling time of the gas is shorter than the viscous time, then a thin disk would form. This is realized for the high gas inflow rate during the processes (such as galaxy mergers) that feed quasars. We start by exploring the structure of thin disks that characterize the high accretion rate of quasars.

Following Shakura & Sunyaev (1973) and Novikov & Thorne (1973), we imagine a planar thin disk of cold gas orbiting a central black hole and wish to describe its structure in polar coordinates (r, ϕ). Each gas element orbits at the local Keplerian velocity $v_\phi = r\Omega = (GM/r)^{1/2}$ and spirals slowly inwards with radial velocity $v_r \ll v_\phi$ as viscous torques transport its angular momentum to the outer part of the disk. The associated viscous stress generates heat, which is radiated away locally from the the disk surface. We assume that the disk is fed steadily and so it manifests a constant mass accretion rate at all radii. Mass conservation implies,

$$\dot{M} = 2\pi r \Sigma v_r = const, \tag{7.8}$$

where $\Sigma(r)$ is the surface mass density of the disk.

In the limit of geometrically thin disk with a scale height $h \ll r$, the hydrodynamic equations decouple in the radial and vertical directions. We start with the radial direction. The Keplerian velocity profile introduces shear which dissipates heat as neighboring fluid elements rub against each other. The concept of shear viscosity can be can be easily understood in the one dimensional example of a uniform gas whose velocity along the y-axis varies linearly with the x coordinate, $V = V_0 + (dV_y/dx)x$. A gas particle moving at the typical thermal speed v traverses a mean-free-path λ along the x-axis before it collides with other particles and shares its y-momentum with them. The y-velocity is different across a distance λ by an amount $\Delta V \sim \lambda dV_y/dx$. Since the flux of particles streaming along the x-axis is $\sim nv$, where n is the gas density, the net flux of y-momentum being transported per unit time, $\sim nvm\Delta V$, is linear in the velocity gradient $\eta dV_y/dx$, with a viscosity coefficient $\eta \sim \rho v\lambda$, where $\rho = mn$ is the mass density of the gas. Since the excess kinetic energy density across a mean-free-path, $\frac{1}{2}\rho(\lambda dV_y/dx)^2$ is dissipated every collision time $\sim (\lambda/v)$, viscosity heats the gas at a rate per unit volume of $\dot{Q} \sim [\eta(dV_y/dx)]^2/\eta$.

Within a Keplerian accretion disk, the flux of ϕ-momentum which is transported

in the positive r-direction is given by the viscous stress $f_\phi = \frac{3}{2}\eta\Omega$, where η is the viscosity coefficient (in $\mathrm{g\ cm^{-1}\ s^{-1}}$) and $\Omega = (GM/r^3)^{1/2}$ is the orbital frequency at a radius r. The viscous stress is expected to be effective down to the ISCO, from where the gas plunges into the black hole on a free fall time. We therefore set the inner boundary of the disk as $r_{\rm ISCO}$, depicted in Figure 7.3. Angular momentum conservation requires that the net rate of its change within a radius r be equal to the viscous torque, namely

$$f_\phi \times (2\pi r \times 2h) \times r = \dot{M}\left[(GMr)^{1/2} - (GMr_{\rm ISCO})^{1/2}\right]. \tag{7.9}$$

The production rate of heat per unit volume by the viscous stress is given by $\dot{Q} = f_\phi^2/\eta$. Substituting f_ϕ and equation (7.9) gives

$$2h\dot{Q} = \frac{3\dot{M}}{4\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.10}$$

This power gives local flux that is radiated vertically from the top and bottom surfaces of the disk,

$$F = \frac{1}{2} \times 2h\dot{Q} = \frac{3\dot{M}}{8\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.11}$$

The total luminosity of the disk is given by

$$L = \int_{r_{\rm ISCO}}^{\infty} 2F \times 2\pi r dr = \frac{1}{2}\frac{GM\dot{M}}{r_{\rm ISCO}}, \tag{7.12}$$

where we have ignored general-relativistic corrections to the dynamics of the gas and the propagation of the radiation it emits.

In the absence of any vertical motion, the momentum balance in the vertical z-direction yields

$$\frac{1}{\rho}\frac{dP}{dz} = -\frac{GM}{r^2}\frac{z}{r}, \tag{7.13}$$

where $z \ll r$ and P and ρ are the gas pressure and density. This equation gives a disk scale height $h \approx c_s/\Omega$ where $c_s \approx (P/\rho)^{1/2}$ is the sound speed.

Because of the short mean-free-path for particles collisions, the particle-level viscosity is negligible in accretion disks. However, such disks are susceptable to the powerful magneto-rotational instability (MRI) that amplifies magnetic turbulence on an orbital time. The origin of the instability can be easly understood by imagining two fluid elements that are threaded by a single magnetic field line and are slightly displaced from each other in the radial direction. The magnetic field acts as a spring owing to its tension. In a Keplerian disc the inner fluid element orbits more rapidly than the outer element, causing the spring to stretch. The inner fluid element is then forced by the spring to slow down, reduce its angular momentum, and therefore move to a lower orbit. The outer fluid element, meanwhile, is forced by the spring to speed up, increase its angular momentum, and therefore move to a higher orbit. The spring tension increases as the two fluid elements separate farther, and eventually the process runs away. The magneto-rotational instability is likely

to develop turbulent eddies in the disk which are much more effective at transporting its angular momentum than particle viscosity. In this case λ and v should be replaced by the typical size and velocity of an eddy. The largest value that these variables can obtain are the scale height h and sound speed c_s in the disk. This implies $f_\phi < (\rho c_s h)\Omega \approx \rho c_s^2 \approx P$. We may then parameterize the viscous stress as some fraction α of its maximum value, $f_\phi = \alpha P$.

The total pressure P in the disk is the sum of the gas pressure $P_{\rm gas} = 2(\rho/m_p)k_B T$, and the radiation pressure, $P_{\rm rad} = \frac{1}{3}aT^4$. We define the fractional contribution of the gas to the total pressure as

$$\beta \equiv \frac{P_{\rm gas}}{P}, \tag{7.14}$$

where $P = P_{\rm rad} + P_{\rm gas}$. In principle, the viscous stress may be limited by the gas pressure only; to reflect this possibility, we write $f_\phi = \alpha P \beta^b$, where b is 0 or 1 if the viscosity scales with the total or just the gas pressure, respectively.

Since the energy of each photon is just its momentum times the speed of light, the radiative energy flux is simply given by the change in the radiation pressure (momentum flux) per photon mean-free-path,

$$F = -c\frac{dP_{\rm rad}}{d\tau}, \tag{7.15}$$

where the optical-depth τ is related to the frequency-averaged (so-called Rosseland-mean) opacity coefficient of the gas, κ,

$$\tau = \int_0^h \kappa\rho dz \approx \frac{1}{2}\kappa\Sigma, \tag{7.16}$$

where $\Sigma = 2h\rho$. For the characteristic mass density ρ and temperature T encountered at the midplane of accretion disks around supermassive black holes, there are two primary sources of opacity: *electron scattering* with

$$\kappa_{\rm es} = \frac{\sigma_{\rm T}}{m_p} = 0.4\ {\rm cm^2\ g^{-1}}, \tag{7.17}$$

and *free-free* absorption with

$$\kappa_{\rm ff} \approx 8 \times 10^{22} {\rm cm^2\ g^{-1}} \left(\frac{\rho}{\rm g\,cm^{-3}}\right)\left(\frac{T}{\rm K}\right)^{-7/2}, \tag{7.18}$$

where we assume a pure hydrogen plasma for simplicity.

It is customary to normalize the accretion rate $\dot{M}$ in the disk relative to the so-called Eddington rate $\dot{M}_E$, which would produce the maximum possible disk luminosity, $L_{\rm Edd}$ (see derivation in equation 7.33 below). When the luminosity approaches the Eddington limit, the disk bloats and h approaches r, violating the thin-disk assumption. We write $\dot{m} = (\dot{M}/\dot{M}_{\rm Edd})$, with $\dot{M}_{\rm Edd} \equiv (L_{\rm Edd}/\epsilon c^2)$, where ϵ is the radiative efficiency for converting rest-mass to radiation near the ISCO.

Based on the above equations, we are now at a position to derive the scaling laws that govern the structure of the disk far away from the ISCO. For this purpose we use the following dimensionless parameters: $r_1 = (r/10R_{\rm Sch})$, $M_8 = (M/10^8 M_\odot)$, $\dot{m}_{-1} = (\dot{m}/0.1)$, $\alpha_{-1} = (\alpha/0.1)$ and $\epsilon_{-1} = (\epsilon/0.1)$.

In local thermodynamic equilibrium, the emergent flux from the surface of the disk (equation 7.11) can be written in terms of the temperature at disk midplane T as $F \approx caT^4/\kappa\Sigma$. The surface temperature of the disk is the roughly,

$$T_d \approx \left(\frac{4F}{a}\right)^{1/4} = 10^5 \text{ K } M_8^{-1/4} \dot{m}_{-1}^{1/4} r_1^{-3/4} \left[1 - \left(\frac{r}{r_{\rm ISCO}}\right)^{1/2}\right]. \quad (7.19)$$

Note that the disk surface temperature rises at low black hole masses and reaches the X-ray regime for stellar-mass black holes. (Non-thermal X-ray emission from a hot corona or a jet can supplement this disk emission.) Stellar mass black holes can therefore be important X-ray sources at high redshifts, especially if they get incorporated into a binary system where they accrete gas from a companion star. In the local Universe, black-hole X-ray binaries come in two flavors, depending on the mass of the companion star: *low-mass x-ray binaries* where a low-mass companion transfers mass owing to the tidal force exerted by the black hole, and *high-mass X-ray binaries (BH-HMXB or micro-quasars)* where the companion is a massive star which could also transfer mass to the black hole through a wind. At redshifts $z > 6$ when the age of the Universe was short, BH-HMXB were probably most important since they are known to produce their X-rays over a short lifetime ($< 10^9$ yr). The cumulative X-ray emission from BH-HMXB is expected to be proportional to the star formation rate. If indeed the early population of stars was tilted towards high masses and binaries were common, BH-HMXB may have been more abundant per star formation rate in high redshift galaxies. As we discuss in other chapters, the X-rays produced by BH-HMXB may have had important observable effects as they catalysed H_2 formation, heated the IGM, and modified the 21-cm signal from neutral hydrogen. Their overall influence was, however, limited: hydrogen could not have been reionized by X-ray sources based on current limits on the unresolved component of the X-ray background. Throughout this chapter, we focus our attention on supermassive black holes, which are brighter and hence easier to detect individually at high redshifts.

For supermassive black holes, the accretion disk can be divided radially into three distinct regions,

1. *Inner region:* where radiation pressure and electron-scattering opacity dominate.
2. *Middle region:* where gas pressure and electron-scattering opacity dominate.
3. *Outer region:* where gas pressure and free-free opacity dominate.

The boundary between regions 1 and 2 is located at the radius

$$r_1 \approx 54\, \alpha_{-1}^{2/21} (\dot{m}_{-1}/\epsilon_{-1})^{16/21} M_8^{2/21} \quad \text{if } b = 1, \quad (7.20)$$

$$58\, \alpha_{-1}^{2/21} (\dot{m}_{-1}/\epsilon_{-1})^{16/21} M_8^{2/21} \quad \text{if } b = 0, \quad (7.21)$$

and the transition radius between regions 2 and 3 is

$$r_1 \approx 4 \times 10^2\, (\dot{m}_{-1}/\epsilon_{-1})^{2/3}. \quad (7.22)$$

The surface density and scale-height of the disk are given by,
Inner region:

$$\Sigma(r) \approx (3\times10^6\,{\rm g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5} \ \text{ if } b=1, \quad (7.23)$$

$$(8\times10^2\,{\rm g\,cm^{-2}})\alpha_{-1}^{-1}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right)^{-1} r_1^{3/2} \quad \text{if } b=0, \quad (7.24)$$

$$h(r) \approx R_{\rm Sch}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right). \quad (7.25)$$

Middle region:

$$\Sigma(r) \approx (3\times10^6\,{\rm g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5}, \quad (7.26)$$

$$h(r) \approx 1.4\times10^{-2} R_S \alpha_{-1}^{-1/10}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right)^{1/5} M_8^{-1/10} r_1^{21/20}. \quad (7.27)$$

Outer region:

$$\Sigma(r) \approx (6\times10^6\,{\rm g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot m_{-1}}{\epsilon_{0.1}}\right)^{7/10} M_8^{1/5} r_1^{-3/4}, \quad (7.28)$$

$$h(r) \approx 10^{-2} R_S \alpha_{-1}^{-1/10}\left(\frac{\dot m_{-1}}{\epsilon_{-1}}\right)^{3/20} M_8^{-1/10} r_1^{9/8}. \quad (7.29)$$

The mid-plane temperature is given by,

$$T_m(r) \approx \left(16\pi^2\right)^{-1/5}\left(\frac{m_p}{k_B\sigma_T}\right)^{1/5} \alpha^{-1/5}\kappa^{1/5}\dot M^{2/5}\Omega^{3/5}\beta^{-(1/5)(b-1)}. \quad (7.30)$$

The above scaling-laws ignore the self-gravity of the disk. This assumption is violated at large radii. The instability of the disk to gravitational fragmentation due to its self-gravity occurs when the so-called Toomre parameter, $Q = (c_s\Omega/\pi G\Sigma)$, drops below unity (see §5.2.3). For the above scaling laws of the outer disk, this occurs at the outer radius,

$$r_1 \approx 2\times10^4 \alpha_{-1}^{28/45}\left(\dot m_{-1}/\epsilon_{-1}\right)^{-22/45} M_8^{52/45}. \quad (7.31)$$

Outside this radius, the disk gas would fragment into stars, and the stars may migrate inwards as the gas accretes onto the black hole. The energy output from stellar winds and supernovae would supplement the viscous heating of the disk and might regulate the disk to have $Q \sim 1$ outside the above boundary. We therefore conclude that star formation will inevitably occur on larger scales, before the gas is driven into the accretion disk that feeds the central black hole. Indeed, the broad emission lines of quasars display very high abundance of heavy elements in the spectra out to arbitrarily high redshifts. Since the total amount of mass in the disk interior to this radius makes only a small fraction of the mass of the supermassive black hole, quasars must be fed by gas that crosses this boundary after being vulnerable to fragmentation.

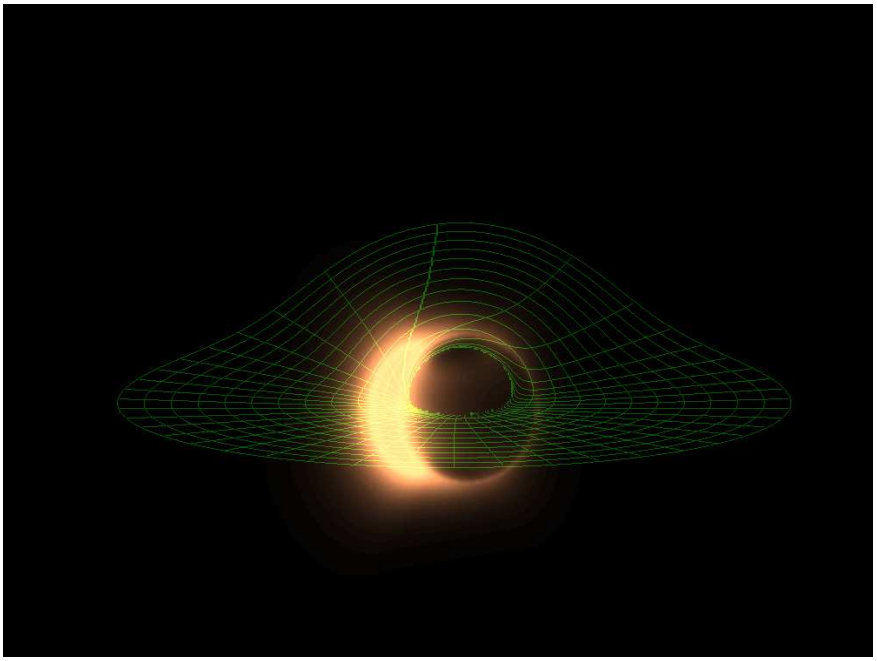

Figure 7.4 Simulated image of an accretion flow around a black hole spinning at half its maximum rate, from a viewing angle of 10° relative to the rotation axis. The coordinate grid in the equatorial plane of the spiraling flow shows how strong lensing around the black hole bends the back of the apparent disk up. The left side of the image is brighter due its rotational motion towards the observer. The bright arcs are generated by gravitational lensing. A dark silhouette appears around the location of the black hole because the light emitted by gas behind it disappears into the horizon and cannot be seen by an observer on the other side. Recently, the technology for observing such an image from the supermassive black holes at the centers of the Milky Way and M87 galaxies has been demonstrated as feasible [Doeleman, S., et al. *Nature* **455**, 78 (2008)]. To obtain the required resolution of tens of micro-arcseconds, signals are being correlated over an array (interferometer) of observatories operating at a millimeter wavelength across the Earth. Figure credit: Broderick, A., & Loeb, A. *Journal of Physics Conf. Ser.* **54**, 448 (2006); *Astrophys. J.* **697** 1164 (2009).

7.2.3 Radiatively Inefficient Accretion Flows

When the accretion rate is considerably lower than its Eddington limit ($\dot{M}/\dot{M}_E < 10^{-2}$), the gas inflow switches to a different mode, called a *Radiatively Inefficient Accretion Flow* (RIAF) or an *Advection Dominated Accretion Flow* (ADAF), in which either the cooling time or the photon diffusion time are much longer than the accretion time of the gas and heat is mostly advected with the gas into the black hole. At the low gas densities and high temperatures characterizing this accretion mode, the Coulomb coupling is weak and the electrons do not heat up to the proton temperature even with the aid of plasma instabilities. Viscosity heats primarily the protons since they carry most of the momentum. The other major heat source, compression of the gas, also heats the protons more effectively than the electrons. As the gas infalls and its density ρ rises, the temperature of each species T increases adiabatically as $T \propto \rho^{\gamma-1}$, where γ is the corresponding adiabatic index. At radii $r < 10^2 r_{\rm Sch}$, the electrons are relativistic with $\gamma = 4/3$ and so their temperature rises inwards with increasing density as $T_e \propto \rho^{1/3}$ while the protons are non-relativistic with $\gamma = 5/3$ and so $T_{\rm p} \propto \rho^{2/3}$, yielding a two-temperature plasma with the protons being much hotter than the electrons. Typical models yield, $T_p \sim 10^{12}\ {\rm K}(r/r_{\rm Sch})^{-1}, T_e \sim \min(T_p, 10^{9-11}\ {\rm K})$. Because the typical sound speed is comparable to the Keplerian speed at each radius, the geometry of the flow is thick – making RIAFs the viscous analogs of Bondi accretions.

Analytic models imply a radial velocity that is a factor of $\sim \alpha$ smaller than the free-fall speed and an accretion time that is a factor of $\sim \alpha$ longer than the free-fall time. However, since the sum of the kinetic and thermal energy of a proton is comparable to its gravitational binding energy, RIAFs are expected to be associated with strong outflows.

The radiative efficiency of RIAFs is smaller than the thin-disk value, ϵ. While the thin-disk value applies to high accretion rates above some critical value, $\dot{m} > \dot{m}_{\rm crit}$, the analytic RIAF models typically admit a radiative efficiency of

$$\frac{L}{\dot{M}c^2} \approx \epsilon \left(\frac{\dot{M}}{\dot{M}_{\rm crit}}\right), \tag{7.32}$$

for $\dot{M} < \dot{M}_{\rm crit}$, with $\dot{M}_{\rm crit}$ in the range of 0.01–0.1. Here $\dot{M}$ is the accretion rate (in Eddington units) near the ISCO, after taking account of the fact that some of the infalling mass at larger radii is lost to outflows. For example, in the nucleus of the Milky Way, massive stars shed $\sim 10^{-3} M_\odot\ {\rm yr}^{-1}$ of mass into the radius of influence of central black hole (SgrA*), but only a tiny fraction $\sim 10^{-5}$ of this mass accretes onto the black hole.

Since at low redshifts mergers are rare and much of the gas in galaxies has already been consumed in making stars, most of the local supermassive black holes are characterized by a very low accretion rate. The resulting low luminosity of these dormant black holes, such as the $4 \times 10^6 M_\odot$ black hole lurking at the center of the Milky Way galaxy, is often described using RIAF/ADAF models. Although this mode of accretion is characterized by a low mass infall rate, it could persist over a period of time that is orders of magnitude longer than the quasar mode discussed earlier and so its contribution to the growth of black holes in galactic nuclei

may not be negligible.

7.3 THE FIRST BLACK HOLES AND QUASARS

A black hole is the end product from the complete gravitational collapse of a material object, such as a massive star. It is surrounded by a horizon from which even light cannot escape. Black holes have the dual virtues of being extraordinarily simple solutions to Einstein's equations of gravity (as they are characterized only by their mass, charge, and spin), but also the most disparate from their Newtonian analogs. In Einstein's theory, black holes represent the ultimate prisons: you can check in, but you can never check out.

Ironically, black hole environments are the brightest objects in the universe. Of course, it is not the black hole that is shining, but rather the surrounding gas is heated by viscously rubbing against itself and shining as it spirals into the black hole like water going down a drain, never to be seen again. The origin of the radiated energy is the release of gravitational binding energy as the gas falls into the deep gravitational potential well of the black hole. As much as tens of percent of the mass of the accreting material can be converted into heat (more than an order of magnitude beyond the maximum efficiency of nuclear fusion). Astrophysical black holes appear in two flavors: stellar-mass black holes that form when massive stars die, and the monstrous super-massive black holes that sit at the center of galaxies, reaching masses of up to 10 billion Suns. The latter type are observed as quasars and active galactic nuclei (AGN). It is by studying these accreting black holes that all of our observational knowledge of black holes has been obtained.

If this material is organized into a thin accretion disk, where the gas can efficiently radiate its released binding energy, then its theoretical modelling is straightforward. Less well understood are radiatively inefficient accretion flows, in which the inflowing gas obtains a thick geometry. It is generally unclear how gas migrates from large radii to near the horizon and how, precisely, it falls into the black hole. We presently have very poor constraints on how magnetic fields embedded and created by the accretion flow are structured, and how that structure affects the observed properties of astrophysical black holes. While it is beginning to be possible to perform computer simulations of the entire accreting region, we are decades away from true *ab initio* calculations, and thus observational input plays a crucial role in deciding between existing models and motivating new ideas.

More embarrassing is our understanding of black hole jets (see images 7.5). These extraordinary exhibitions of the power of black holes are moving at nearly the speed of light and involve narrowly collimated outflows whose base has a size comparable to the solar system, while their front reaches scales comparable to the distance between galaxies. Unresolved issues are as basic as what jets are made of (whether electrons and protons or electrons and positrons, or primarily electromagnetic fields) and how they are accelerated in the first place. Both of these rest critically on the role of the black hole spin in the jet-launching process.

A quasar is a point-like ("quasi-stellar") bright source at the center of a galaxy. There are many lines of evidence indicating that a quasar involves a supermassive

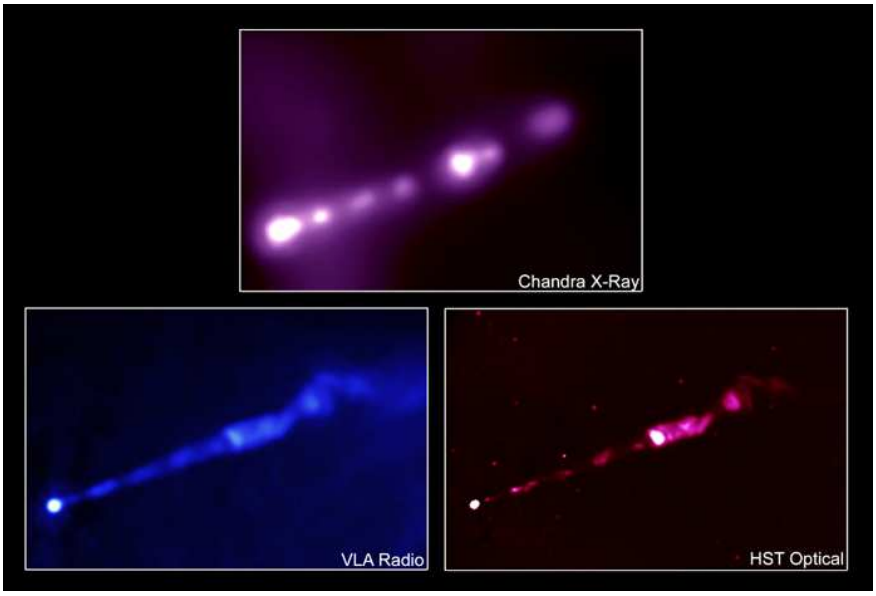

Figure 7.5 Multi-wavelength images of the highly collimated jet emanating from the supermassive black hole at the center of the giant elliptical galaxy M87. The X-ray image (top) was obtained with the Chandra X-ray satellite, the radio image (bottom left) was obtained with the Very Large Array (VLA), and the optical image (bottom right) was obtained with the Hubble Space Telescope (HST).

black hole, weighting up to ten billion Suns, which is accreting gas from the core of its host galaxy. The supply of large quantities of fresh gas is often triggered by a merger between two galaxies. The infalling gas heats up as it spirals towards the black hole and dissipates its rotational energy through viscosity. The gas is expected to be drifting inwards in an accretion disk whose inner "drain" has the radius of the ISCO, according to Einstein's theory of gravity. Interior to the ISCO, the gas plunges into the black hole in such a short time that it has no opportunity to radiate most of its thermal energy. However, as mentioned in §7.1 the fraction of the rest mass of the gas which gets radiated away just outside the ISCO is high, ranging between 5.7% for a non-spinning black hole to 42.3% for a maximally-spinning black hole (see Figure 7.3). This "radiative efficiency" is far greater than the mass-energy conversion efficiency provided by nuclear fusion in stars, which is $< 0.7\%$.

Quasar activity is observed in a small fraction of all galaxies at any cosmic epoch. Mammoth black holes weighing more than a billion solar masses were discovered at redshifts as high as $z \sim 6.5$, less than a billion years after the Big Bang. *If massive black holes grow at early cosmic times, should their remnants be around us today?* Indeed, searches for black holes in local galaxies have found that every galaxy with a stellar spheroid harbors a supermassive black hole at its center. This implies that quasars are rare simply because their activity is short-lived. Moreover, there appears to be a tight correlation between the black hole mass and the gravitational potential-well depth of their host spheroids of stars (as measured by the velocity dispersion of these stars). This suggests that the black holes grow up to the point where the heat they deposit into their environment or the piston effect from their winds prevent additional gas from feeding them further. The situation is similar to a baby who gets more energetic as he eats more at the dinner table, until his hyper-activity is so intense that he pushes the food off the table and cannot eat any more. This *principle of self-regulation* explains why quasars are short lived and why the final black hole mass is dictated by the depth of the potential in which the gas feeding it resides. It is also possible the feedback from star formation in the vicinity of the black holes affects or controls their self-regulation. Most black holes today are dormant or "starved" because the gas around them was mostly used up in making the stars, or because their activity heated or pushed it away a long time ago.

What seeded the formation of supermassive black holes only a billion years after the Big Bang? We know how to make a black hole out of a massive star. When the star ends its life, it stops producing sufficient energy to hold itself against its own gravity, and its core collapses to make a black hole. Long before evidence for black holes was observed, this process leading to their existence was understood theoretically by Robert Oppenheimer and Hartland Snyder in 1937. However, growing a supermassive black hole is more difficult. There is a maximum luminosity at which the environment of a black hole of mass $M_{\rm BH}$ may shine and still accrete gas.[i] This

[i]Whereas the gravitational force acts mostly on the protons, the radiation force acts primarily on the electrons. These two species are tied together by a global electric field, so that the entire "plasma" (ionized gas) behaves as a single quasi-neutral fluid which is subject to both forces. Under similar circumstances, electrons are confined to the Sun by an electric potential of about a kilo-Volt (corre-

Eddington luminosity, L_E, was derived in equation (5.27) by balancing the inward force of gravity on each proton by the outward radiation force on its companion electron (which is the momentum flux carried by the radiation times the scattering cross-section of the electron) at a distance r:

$$\frac{GM_{\rm BH}m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{7.33}$$

where m_p is the proton mass and $\sigma_T = 0.67 \times 10^{-24}$ cm^2 is the cross-section for scattering a photon by an electron. Interestingly, the limiting luminosity is independent of radius in the Newtonian regime. Since the Eddington luminosity represents an exact balance between gravity and radiation forces, it actually equals the luminosity of massive stars which are held at rest against gravity by radiation pressure, as described by equation (7.34). This limit is formally valid in a spherical geometry, and exceptions to it were conjectured for other accretion geometries over the years. But, remarkably, observed quasars for which black hole masses can be measured by independent methods appear to respect this limit. Substituting all constants, the Eddington luminosity is given by,

$$L_E = 1.3 \times 10^{44} \left(\frac{M_{\rm BH}}{10^6 M_\odot}\right) \ {\rm erg\ s^{-1}}, \tag{7.34}$$

Interestingly, the scattering cross section per unit mass for UV radiation on dust is larger by two orders of magnitude than σ_T/m_p. Although dust is destroyed within $\sim 10^4 GM_{\rm BH}/c^2$ by the strong illumination from an Eddington-limited quasar, it should survive at larger distances. Hence, the radiation pressure on dust would exceed the gravitational force towards the black hole and drive powerful outflows. Spectral lines could be even more effective than dust in their coupling to radiation. The integral of the absorption cross-section of a spectral line over frequency,

$$\int \sigma(\nu) d\nu = f_{12}\left(\frac{\pi e^2}{m_e c}\right), \tag{7.35}$$

is typically orders of magnitude larger than $\sigma_T \nu_{21}$ where ν_{21} is the transition frequency and f_{12} is the absorption oscillator strength. For example, the Lyman-α transition of hydrogen, for which $f_{12} = 0.416$, provides an average cross-section which is seven orders of magnitude larger than σ_T when averaged over a frequency band as wide as the resonant frequency itself. Therefore, lines could be even more effective at driving outflows in the outer parts of quasar environments.

As discussed before, the total luminosity from gas accreting onto a black hole, L, can be written as some radiative efficiency ϵ times the mass accretion rate $\dot{M}$,

$$L = \epsilon \dot{M} c^2, \tag{7.36}$$

---

sponding to a total charge of ~ 75 Coulombs). The opposite electric forces per unit volume acting on electrons and ions in the Sun cancel out so that the total pressure force is exactly balanced by gravity, as for a neutral fluid. An electric potential of 1-10 kilo-Volts also binds electrons to clusters of galaxies (where the thermal velocities of these electrons, $\sim 0.1c$, are well in excess of the escape speed from the gravitational potential). For a general discussion, see Loeb, A. *Phys. Rev.* **D37**, 3484 (1988).

with the black hole accreting the non-radiated component, $\dot{M}_{\rm BH} = (1-\epsilon)\dot{M}$. The equation that governs the growth of the black hole mass is then

$$\dot{M}_{\rm BH} = \frac{M_{\rm BH}}{t_E}, \tag{7.37}$$

where (after substituting all fundamental constants),

$$t_E = 4 \times 10^7 {\rm years} \left(\frac{\epsilon/(1-\epsilon)}{10\%}\right) \left(\frac{L}{L_E}\right)^{-1}. \tag{7.38}$$

We therefore find that as long as fuel is amply supplied, the black hole mass grows exponentially in time, $M_{\rm BH} \propto \exp\{t/t_E\}$, with an e-folding time t_E. Since the growth time in equation (7.38) is significantly shorter than the $\sim 10^9$ years corresponding to the age of the Universe at a redshift $z \sim 6$ – where black holes with a mass $\sim 10^9 M_\odot$ are found, one might naively conclude that there is plenty of time to grow the observed black hole masses from small seeds. For example, a seed black hole from a Population III star of $100 M_\odot$ can grow in less than a billion years up to $\sim 10^9 M_\odot$ for $\epsilon \sim 10\%$ and $L \sim L_E$. However, the intervention of various processes makes it unlikely that a stellar mass seed will be able to accrete continuously at its Eddington limit with no interruption.

For example, mergers are very common in the early Universe. Every time two gas-rich galaxies come together, their black holes are likely to coalesce. The coalescence is initially triggered by "dynamical friction" on the surrounding gas and stars, and is completed – when the binary gets tight – as a result of the emission of gravitational radiation. The existence of gravitational waves is a generic prediction of Einstein's theory of gravity. They represent ripples in space-time generated by the motion of the two black holes as they move around their common center of mass in a tight binary. The energy carried by the waves is taken away from the kinetic energy of the binary, which therefore gets tighter with time. Computer simulations reveal that when two black holes with unequal masses merge to make a single black hole, the remnant gets a kick due to the non-isotropic emission of gravitational radiation at the final plunge.[ii] This kick was calculated recently using advanced computer codes that solve Einstein's equations (a task that was plagued for decades with numerical instabilities). The typical kick velocity is hundreds of kilometer per second (and up to ten times more for special spin orientations), bigger than the escape speed from the first dwarf galaxies. This implies that continuous accretion was likely punctuated by black hole ejection events, forcing the merged dwarf galaxy to grow a new black hole seed from scratch.[iii]

If continuous feeding is halted, or if the black hole is temporarily removed from the center of its host galaxy, then one is driven to the conclusion that the black

[ii] The gravitational waves from black hole mergers at high redshifts could in principle be detected by a proposed space-based mission called the *Laser Interferometer Space Antenna* (LISA). For more details, see http://lisa.nasa.gov/, and, for example, Wyithe, J. S. B., & Loeb, A. *Astrophys. J.* **590**, 691 (2003).

[iii] These black hole recoils might have left observable signatures in the local Universe. For example, the halo of the Milky Way galaxy may include hundreds of freely-floating ejected black holes with compact star clusters around them, representing relics of the early mergers that assembled the Milky Way out of its original building blocks of dwarf galaxies (O'Leary, R. & Loeb, A. *Mon. Not. R. Astron. Soc.* **395**, 781 (2009)).

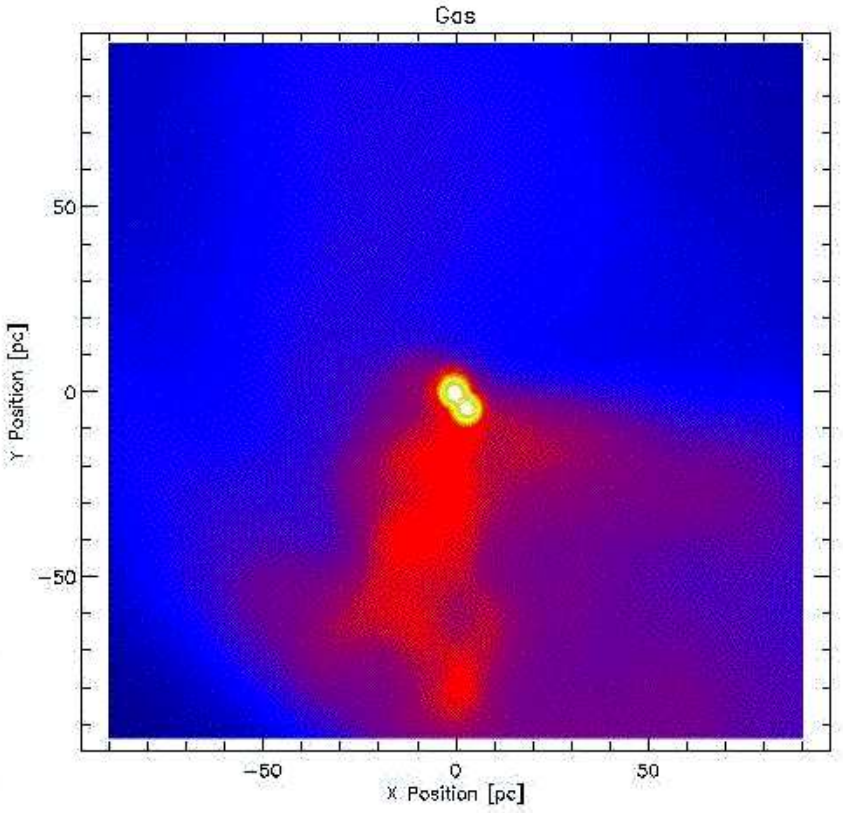

Figure 7.6 Numerical simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2. The image shows a snapshot of the gas density distribution 500 million years after the Big Bang, indicating the formation of two compact objects near the center of the galaxy with masses of $2.2 \times 10^6 M_\odot$ and $3.1 \times 10^6 M_\odot$, respectively, and radii < 1 pc. Sub-fragmentation into lower mass clumps is inhibited because hydrogen atoms cannot cool the gas significantly below its initial temperature. These circumstances lead to the formation of supermassive stars that inevitably collapse to make massive seeds of supermassive black holes. The simulated box size is 200 pc on a side. Figure credit: Bromm, V. & Loeb, A. *Astrophys. J.* **596**, 34 (2003).

hole seeds must have started more massive than $\sim 100 M_\odot$. More massive seeds may originate from supermassive stars. *Is it possible to make such stars in early galaxies?* Yes, it is. Numerical simulations indicate that stars weighing up to a million Suns could have formed at the centers of early dwarf galaxies which were barely able to cool their gas through transitions of atomic hydrogen, having $T_{\rm vir} \sim 10^4$K and no H_2 molecules. Such systems have a total mass that is several orders of magnitude higher than the earliest Jeans-mass condensations discussed in §3.2. In both cases, the gas lacks the ability to cool well below $T_{\rm vir}$, and so it fragments into one or two major clumps. The simulation shown in Figure 7.6 results in clumps of several million solar masses, which inevitably end up as massive black holes. The existence of such seeds would have given a jump start to the black hole growth process.

Supermassive stars, defined as hydrostatic configurations with masses 10^3–$10^8 M_\odot$, have not been observed as of yet. Theoretically, they are expected to be supported almost entirely by radiation pressure and hence their luminosity equals the Eddington limit, $L = 1.3 \times 10^{44} (M_\star/10^6 M_\odot)\ {\rm ergs\,s^{-1}}$. Supermassive stars steadily contract and convert their gravitational binding energy to radiation with a total lifetime $< 10^6$ yr before they collapse to a black hole. First we show that the envelope of such stars must be convective.

The condition for convective instability is that that the star exhibits a negative entropy gradient. This follows from the fact that convective eddies which are hotter and rarefied relative to their environment tend to rise towards the star's surface and decrease their density adiabatically (at constant entropy) in pressure equilibrium with their environment. If the background entropy decreases as the eddies rise, then they become even more rarefied relative to their environment (lower ambient entropy at the same pressure implies higher ambient density) and continue to rise even further, hence leading to an instability. The energy transport by convective eddies drives the star to a state of marginal stability, namely nearly uniform entropy. Let us first show that in the absence of convection, a supermassive star will tend to develop a negative entropy gradient as it radiates away its energy.

The entropy of each electron-proton fluid element in a supermassive star is changing according to the local radiative heat flux $\mathbf{F}$ at a rate,

$$\frac{2T_g}{m_p}\frac{\partial s}{\partial t} = -\frac{1}{\rho}\nabla\cdot\mathbf{F}, \tag{7.39}$$

where ρ, T_g and s are the mass density, temperature and specific entropy of the element. If the opacity is dominated by Thomson scattering, then the local radiative heat flux $\mathbf{F}$ is related to the radiation pressure gradient by,

$$\mathbf{F} = \frac{-m_p}{\sigma_T}\frac{1}{\rho}\nabla p_\gamma. \tag{7.40}$$

Ignoring gas pressure and rotation, the hydrostatic equilibrium equation is simply,

$$\frac{1}{\rho}\nabla p_\gamma = \mathbf{g}, \tag{7.41}$$

where the gravitational field $\mathbf{g}$ obeys Poisson's equation,

$$\nabla\cdot\mathbf{g} = -4\pi G\rho. \tag{7.42}$$

Combining equations (7.40)-(7.42), we find that the right-hand side of equation (7.39) is constant,

$$\frac{1}{\rho}\nabla\cdot\mathbf{F}=\frac{4\pi G m_p}{\sigma_T}=const. \tag{7.43}$$

Therefore, the gradient of equation (7.39) gives,

$$\frac{\partial}{\partial t}\nabla s=\frac{2\pi G m_p^2}{\sigma_T}\frac{\nabla T_g}{T_g^2}<0. \tag{7.44}$$

The radial temperature gradient is negative since heat flows out of the star, implying that the star will develop a negative entropy gradient and become convectively unstable. This result holds also for a rotating star, as long as the rotation period is much longer than $(G\rho)^{-1/2}=1.1\ \mathrm{hr}(\rho/1\ \mathrm{g\ cm^{-3}})^{-1/2}$.

The nearly uniform entropy established by convection makes the structure of supermassive stars simple (equivalent to a so-called polytrope with an index $n=3$) with a unique relation between their central temperature T_c and central density ρ_c,

$$T_c=2\times10^6\ \mathrm{K}\left(\frac{\rho_c}{1\ g\ cm^{-3}}\right)^{1/3}\left(\frac{M}{10^6 M_\odot}\right)^{1/6}. \tag{7.45}$$

Hence, nuclear reactions are insignificant in metal-poor stars with masses $M_\star > 10^5 M_\odot$. General relativistic corrections make the star unstable to direct collapse to a black hole as soon as its radius contracts to a value,

$$R_\star < R_{\rm crit}=1.59\times10^3\left(\frac{M_\star}{10^6 M_\odot}\right)^{1/2}\left(\frac{GM_\star}{c^2}\right). \tag{7.46}$$

Rotation can stabilize supermassive stars to smaller radii, but even rotating stars are expected to evetually collapse to a black hole after shedding their angular momentum through a wind. If the supermassive star is made of pre-enriched gas, then powerful winds will inevitably be driven at its surface where the opacity due to lines from heavy elements far exceeds the Thomson value, making the outward radiation force stronger than gravity.

We note that the infall of a sufficiently dense, optically-thick spherical envelope of gas cannot be prevented by radiation pressure even if the radiation production rate exceeds the Eddington limit near the center. To see this, let us consider a gas shell falling inwards with a velocity $v_{\rm in}(r)$ at a radius r. If the outward diffusion time of photons through the gas, $t_{\rm diff}\sim\tau r/c$, exceeds the infall time, $t_{\rm in}\sim r/v_{\rm in}$, then the radiation will be dragged by the infalling gas into the black hole. Even though the radiation is always diffusing outwards in the local rest-frame of the gas, it actually moves inwards in the black hole frame of reference when $t_{\rm diff}>t_{\rm in}$. In that regime, the radiation will never be able to counteract the collapse of gas shells that are farther out. Here $\tau\sim(\sigma_T/m_p)\rho r$ is the shell's optical depth to Thomson scattering. Expressing the mass accretion rate as,

$$\dot{M}=4\pi\rho r^2 v_{\rm in}, \tag{7.47}$$

we find that $t_{\rm diff}>t_{\rm in}$ if

$$\frac{\dot{M}}{\dot{M}_{\rm E}}>\epsilon\left(\frac{r}{GM/c^2}\right), \tag{7.48}$$

where M is the mass interior to radius r and $\dot{M}_{\rm E} = L_{\rm E}/\epsilon c^2$ is the mass accretion rate that produces the Eddington luminosity $L_{\rm E} = (4\pi G M m_p c/\sigma_T)$ for a radiative efficiency ϵ. We therefore conclude that as long as the mass infall rate is sufficiently high, the Eddington limit will not apply because of photon trapping. Super-Eddington accretion can therefore grow a seed black hole rapidly, as long as the blanket of infalling gas advects the radiation inwards as it accretes onto the black hole. This "obscured" mode of black hole accretion (which is hidden from view for observers) could be particularly important at high redshifts when the gas density and infall rate onto galaxies obtain their highest values.

The nuclear black holes in galaxies are believed to be fed with gas in episodic events of gas accretion triggered by mergers of galaxies. The energy released by the accreting gas during these episodes could easily unbind the gas reservoir from the host galaxy and suppress star formation within it. If so, nuclear black holes regulate their own growth by expelling the gas that feeds them. In so doing, they also shape the stellar content of their host galaxy. This may explain the observed tight correlations between the mass of central black holes in present-day galaxies and the velocity dispersion $\sigma_\star$ or luminosity $L_{\rm sp}$ of their host spheroids of stars (namely, $M_{\rm BH} \propto \sigma_\star^4$ or $M_{\rm BH} \propto L_{\rm sp}$). Since the mass of a galaxy at a given redshift scales with its virial velocity as $M \propto V_c^3$ in equation (3.32), the binding energy of galactic gas is expected to scale as $MV_c^2 \propto V_c^5$ while the momentum required to kick the gas out of its host would scale as $MV_c \propto V_c^4$. Both scalings can be tuned to explain the observed correlations between black hole masses and the spheroid velocity dispersion of their host galaxies, shown in Figure **??**. Star formation inevitably precedes black hole fueling, since the outer region of the accretion flows that feed nuclear black holes is typically unstable to fragmentation.

The feedback regulated growth explains why quasars may shine much brighter than their host galaxies. A typical star like the Sun emits a luminosity, $L_\odot = 4\times10^{33}\,{\rm erg\,s^{-1}}$ which can also be written as a fraction $\sim 3\times10^{-5}$ of its Eddington luminosity $L_E = 1.4\times10^{38}\,{\rm erg\,s^{-1}}$. Black holes grow up to a fraction $\sim 10^{-3}$ of the stellar mass of their spheroid. When they shine close to their Eddington limit, they may therefore outshine their host galaxy by up to a factor of $\sim (10^{-3}/3\times 10^{-5})$, namely 1–2 orders of magnitude. The factor is smaller during short starburst episodes which are dominated by massive stars with larger Eddington fractions.

The inflow of cold gas towards galaxy centers during the growth phase of their black holes would naturally be accompanied by a burst of star formation. The fraction of gas not consumed by stars or ejected by supernova-driven winds will continue to feed the black hole. It is therefore not surprising that quasar and starburst activities co-exist in ultra-luminous galaxies, and that all quasars show strong spectral lines of heavy elements. Similarly to the above-mentioned prescription for modeling galaxies, it is possible to "dress up" the mass distribution of halos in Figure 3.4 with quasar luminosities (related to L_E, which is a prescribed function of M based on the observed $M_{\rm BH}$–$\sigma_\star$ relation) and a duty cycle (related to t_E or the dynamical time of the host galactic disk), and find the evolution of the quasar population over redshift. This simple approach can be tuned to give good agreement with the data on the quasar luminosity function shown in Figure 7.1. To get reasonable agreement with observations, one needs to assume that quasars deposit

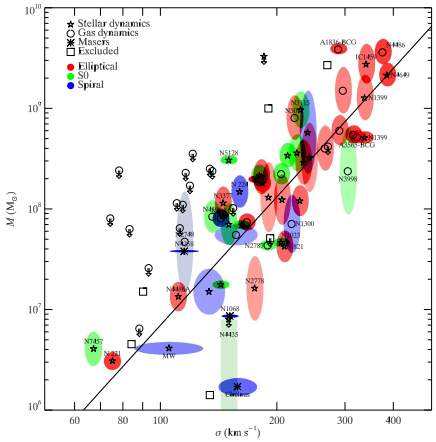

Figure 7.7 Dynamical measurements of the correlation between supermassive black hole mass ($M_{\rm BH}$) and velocity dispersion of stars in the spheroid of its host galaxy (σ). The symbol indicates the method of black hole mass measurement: dynamics of stars (*pentagrams*), dynamics of gas (*circles)*, dynamics of maser sites (*asterisks*). Arrows indicate 3σ upper limits to black hole mass. The shade of the error ellipse indicates the Hubble type of the host galaxy: elliptical, S0, or spiral. The line is the best fit relation to the full sample: $M_{\rm BH} = 10^{8.12}\ M_\odot (\sigma/200\ {\rm km\ s^{-1}})^{4.24}$. The mass uncertainty for NGC 4258 has been plotted much larger than its actual value so that it will show on this plot. Figure credit: K. Gültekin, et al., *Astrophys. J.* **698**, 198 (2009).

$\sim 5\%$ of their Eddington luminosity in the ISM of their host galaxy. The coupling mechanism is unknown and could be related to either the bright radiation or fast outflows that are known to be produced by quasars.

The early growth of massive black holes led to the supermassive black holes observed today. In our own Milky Way galaxy, stars are observed to zoom around the Galactic center at speeds of up to ten thousand kilometers per second, owing to the strong gravitational acceleration near the central black hole. But closer-in observations are forthcoming. Existing technology should soon be able to image the silhouette of the supermassive black holes in the Milky Way and M87 galaxies directly (see Figure 7.4).

7.4 BLACK HOLE BINARIES

Nearly all nearby galactic spheroids are observed to host a nuclear black hole. Therefore, the hierarchical buildup of galaxies through mergers must generically produce black hole binaries. Such binaries tighten through dynamical friction on the background gas and stars, and ultimately coalesce through the emission of gravitational radiation.

In making a tight binary from a merger of separate galaxies, the mass ratio of two black holes cannot be too extreme. A satellite of mass $M_{\rm sat}$ in a circular orbit at the virial radius of a halo of mass $M_{\rm halo}$ would sink to the center on a dynamical friction time of $\sim 0.1 t_H (M_{\rm halo}/M_{\rm sat})$, where t_H is the Hubble time. If the orbit is eccentric with an angular momentum that is a fraction ε of a circular orbit with the same energy, then the sinking time reduces by a factor of $\sim \varepsilon^{0.4}$. Therefore, mostly massive satellites with $M_{\rm sat} > 0.1 M_{\rm halo}$ bring their supermassive black holes to the center of their host halos during the age of the Universe.

As a satellite galaxy sinks, its outer envelope of dark matter and stars is stripped by tidal forces. The stripping is effective down to a radius inside of which the mean mass density of the satellite is comparable to the ambient density of the host galaxy. Eventually, the two black holes are stripped down to the cores of their original galaxies and are surrounded by a circumbinary envelope of stars and gas. As long as the binary is not too tight, the reservoir of stars within the binary orbit can absorb the orbital binding energy of the binary and allow it to shrink. However, when the orbital velocity starts to exceed the local velocity dispersion of stars, a star impinging on the binary would typically be expelled from the galactic nucleus at a high speed. This happens at the so-called the "hardening radius" of the binary,

$$a_{\rm hard} \approx 0.1 \frac{q}{(1+q)^2} M_6 \left(\frac{\sigma_\star}{100\ {\rm km\ s^{-1}}} \right)^{-2} {\rm pc}, \tag{7.49}$$

at which the binding energy per unit mass of the binary exceeds $\frac{3}{2}\sigma^2$, where σ is the velocity dispersion of the stars before the binary tightened. Here, $M \equiv (M_1 + M_2)$, $M_6 = (M/10^6 M_\odot)$, where M_1 and M_2 are the masses of the two black holes, $q = M_1/M_2$ is their mass ratio, and $\mu = M_1 M_2/(M_1 + M_2)$ is the reduced mass of the binary.

A hard binary will continue to tighten only by expelling stars which cross its

orbit and so unless the lost stars are replenished by new stars which are scattered into an orbit that crosses the binary (through dynamical relaxation processes in the surrounding galaxy) the binary would stall. This "final parsec problem" is circumvented if gas streams into the binary from a circumbinary disk. Indeed, the tidal torques generated during a merger extract angular momentum from any associated cold gas and concentrate the gas near the center of the merger remnant, where its accretion often results in a bright quasar.

If the two black holes are in a circular orbit of radius $a < a_{\rm hard}$ around each other, their respective distances from the center of mass are $a_i = (\mu/M_i)a$ ($i = 1, 2$). We define the parameter $\zeta = 4\mu/(M_1+M_2)$, which equals unity if $M_1 = M_2$ and is smaller otherwise. The orbital period is given by

$$P = 2\pi(GM/a^3)^{-1/2} = 1.72 \times 10^{-2}\ {\rm yr}\ a_{14}^{3/2} M_6^{-1/2}, \tag{7.50}$$

where, $a_{14} \equiv (a/10^{14}\ {\rm cm})$. The angular momentum of the binary can be expressed in terms of the absolute values of the velocities of its members v_1 and v_2 as $J = \Sigma_{i=1,2} M_i v_i a_i = \mu v a$, where the relative orbital speed is

$$v = v_1 + v_2 = (2\pi a/P) = 1.15 \times 10^4\ {\rm km\ s^{-1}} M_6^{1/2} a_{14}^{-1/2}\ . \tag{7.51}$$

In gas-rich mergers, the rate of inspiral slows down as soon as the gas mass interior to the binary orbit is smaller than μ and the enclosed gas mass is no longer sufficient for carrying away the entire orbital angular momentum of the binary, J. Subsequently, momentum conservation requires that fresh gas will steadily flow towards the binary orbit in order for it to shrink. The binary tightens by expelling gas out of a region twice as large as its orbit (similarly to a "blender" opening a hollow gap) and by torquing the surrounding disk through spiral arms. Fresh gas re-enters the region of the binary as a result of turbulent transport of angular momentum in the surrounding disk. Since the expelled gas carries a specific angular momentum of $\sim va$, the coalescence time of the binary is inversely proportional to the supply rate of fresh gas into the binary region. In a steady state, the mass supply rate of gas that extracts angular momentum from the binary, $\dot{M}$, is proportional to the accretion rate of the surrounding gas disk. Given that a fraction of the mass that enters the central gap accretes onto the BHs and fuels quasar activity, it is appropriate to express $\dot{M}$ in Eddington units $\dot{\mathcal{M}} \equiv \dot{M}/\dot{M}_E$, corresponding to the accretion rate required to power the limiting Eddington luminosity with a radiative efficiency of 10%, $\dot{M}_E = 0.023 M_\odot\ {\rm yr}^{-1} M_6$. We then find,

$$t_{\rm gas} \approx (J/\dot{M}va) = \mu/\dot{M} = 1.1 \times 10^7\ {\rm yr}\ \zeta \dot{\mathcal{M}}^{-1}. \tag{7.52}$$

For a steady $\dot{\mathcal{M}}$, the binary spends equal amounts of time per log a until GWs start to dominate its loss of angular momentum.

The coalescence timescale due to GW emission is given by,

$$t_{\rm GW} = \frac{5}{256}\frac{c^5 a^4}{G^3 M^2 \mu} = 2.53 \times 10^3\ {\rm yr}\ \frac{a_{14}^4}{\zeta M_6^3}. \tag{7.53}$$

By setting $t_{\rm GW} = t_{\rm gas}$ we can solve for the orbital speed, period, and separation at which GWs take over,

$$v_{\rm GW} = 4.05 \times 10^3\ {\rm km\ s^{-1}}\ \zeta^{-1/4} (\dot{\mathcal{M}} M_6)^{1/8}\ ; \tag{7.54}$$

$$P_{\rm GW} = 0.4\ {\rm yr}\ \zeta^{3/4} M_6^{5/8} \dot{\mathcal{M}}^{-3/8}\ ; \tag{7.55}$$

$$a_{\rm GW} = 2.6 \times 10^{-4}\ {\rm pc}\ \zeta^{1/2} M_6^{3/4} \dot{\mathcal{M}}^{-1/4}. \tag{7.56}$$

For a binary redshift z, the observed period is $(1+z)P_{\rm GW}$. The orbital speed at which GWs take over is very weakly dependent on the supply rate of gas, $v_{\rm GW} \propto \dot{M}^{1/8}$. It generically corresponds to an orbital separation of order $\sim 10^3$ Schwarzschild radii ($2GM/c^2$). The probability of finding binaries deeper in the GW-dominated regime, $\mathcal{P} \propto t_{\rm GW}$, diminishes rapidly at increasing orbital speeds, with $\mathcal{P} = \mathcal{P}_{\rm GW}(v/v_{\rm GW})^{-8}$.

Black hole binaries can be identified visually or spectroscopically. At large separations the cores of the merging galaxies can be easily identified as separate entities. If both black holes are active simultaneously, then the angular separation between the brightness centroids can in principle be resolved at X-ray, optical, infrared, or radio wavelengths. The UV illumination by a quasar usually produces narrow lines from gas clouds at kpc distances within its host galaxy or broad lines from denser gas clouds at sub-pc distances from it. Therefore the existence of a binary can be inferred from various spectroscopic offsets: *(i)* between two sets of narrow lines if the galaxies are separated by more than a few kpc and both have quasar activity at the same time; *(ii)* between the narrow emission lines of the gas and the absorption lines of the stars due to the tidal interaction between the galaxies at a multi-kpc separation; *(iii)* between narrow lines and broad lines if the black hole binary separation is between the kpc and pc scales. The last offset signature can also be produced by a single quasar which gets kicked out of the center of its host galaxy while carrying the broad-line region with it. Such a kick could be produced either by the anisotropic emission of gravitational waves during the coalescence of a binary (producing a recoil of up to $\sim 200\ {\rm km\ s^{-1}}$ in a merger of non-spinning black holes, and up to $\sim 4,000\ {\rm km\ s^{-1}}$ for special spin orientation), or from triple black hole systems that form when a third black hole is added to a galaxy center before the binary there had coalesced. Aside from testing general relativity in the strong field limit, fast recoils have an important feedback effect in forcing a fresh start for the growth of black holes in small galaxies at high redshifts. These early recoils may have also left a fossil signatures in the local Universe: for example, the hierarchical formation of the Milky-Way may have left recoiled black holes floating in its halo, which are detectable through the compact star clusters that remain bound to these intermediate-mass black holes following their ejection from their host dwarf galaxies at high redshifts.

Chapter Eight

The Reionization of Cosmic Hydrogen by the First Galaxies

8.1 IONIZATION SCARS BY THE FIRST STARS

The CMB indicates that hydrogen atoms formed 400 thousand years after the Big Bang, as soon as the gas cooled below 3,000K as a result of cosmological expansion. On the other hand, observations of the CMB as well as of the spectra of early galaxies, quasars, and gamma-ray bursts indicate that less than a billion years later the same gas underwent a wrenching transition from atoms back to their constituent protons and electrons in a process known as **reionization**. More specifically, the $z \sim 6$ Lyman-α forest shows that the IGM is highly-ionized at this time, though there are possible hints from other methods that some large neutral hydrogen regions persist at these early times, which suggests that we may not need to go to much higher redshifts to begin to see the epoch of reionization. Moreover, CMB polarization studies demand that the universe could not have fully reionized earlier than an age of 300 million years. It is intriguing that the inferred reionization epoch coincides with the appearance of the first galaxies, which inevitably produced ionizing radiation. *How was the primordial gas transformed to an ionized state by the first galaxies within merely hundreds of million of years?*

We begin this chapter by addressing this question using our tools describing the formation and evolution of galaxies during the cosmic dawn. The course of reionization can be determined by counting photons from all galaxies as a function of time. Both stars and black holes contribute ionizing photons, but the early Universe is dominated by small galaxies which, in the local universe, have disproportionately small central black holes. In fact, bright quasars are known to be extremely rare at $z > 6$, so we will generally focus on stellar models as a fiducial case.

Because stellar ionizing photons are only slightly more energetic than the 13.6 eV ionization threshold of hydrogen, they are absorbed efficiently once they reach a region with substantial neutral hydrogen. This makes the IGM during reionization a two-phase medium, characterized by highly ionized zones separated from the neutral sea of gas by sharp ionization fronts. While the redshift at which reionization ended only constrains the overall cosmic efficiency for producing ionizing photons, a detailed picture of these structures in progress will teach us a great deal about the population of the first galaxies that produced this cosmic phase transition.

8.2 PROPAGATION OF IONIZATION FRONTS

The simplest reionization problem is to consider how a single, isolated galaxy ionizes its surroundings. The formation of H II regions, or ionized bubbles, around galaxies is the fundamental process that drives reionization, although in practice these galaxies are only isolated in the very earliest phases of reionization. Our first goal is to model this problem of an isolated expanding H II region.

Let us consider, for simplicity, a spherical ionized volume V, separated from the surrounding neutral gas by a sharp ionization front. In the absence of recombinations, eachhydrogen atom in the IGM would only have to be ionized once, and the ionized physical volume V_p would simply be determined by

$$\bar{n}_H V_p = N_\gamma \ , \tag{8.1}$$

where $\bar{n}_H$ is the mean number density of hydrogen and N_γ is the total number of ionizing photons produced by the source.

The size of the resulting H II region depends on the halo which produces it. Let us consider a halo of total mass M and baryon fraction Ω_b/Ω_m. To derive a rough estimate, we assume that baryons are incorporated into stars with an efficiency $f_\star$ and that the escape fraction for the resulting ionizing radiation is $f_{\rm esc}$. We also let $N_{\rm ion}$ be the number of ionizing photons per baryon inside stars; this is $\sim 4,000$ for Population II stars with a "normal" IMF. We finally introduce a parameter $A_{\rm He} = 4/(4-3Y_p) = 1.22$, where Y_p is the mass fraction of helium, as a correction factor to convert the number of ionizing photons to the number of ionized hydrogen atoms (assuming that helium is singly ionized as well). At least in our simple model, so far as the IGM is concerned all these parameters are completely degenerate and determine the overall ionizing efficiency, which we will call ζ,

$$\zeta = A_{\rm He} f_\star f_{\rm esc} N_{\rm ion}. \tag{8.2}$$

If we neglect recombinations, then we obtain the maximum comoving radius of the region which the halo of mass M can ionize,

$$\begin{aligned} r_{\rm max} = \left(\frac{3}{4\pi} \frac{N_\gamma}{\bar{n}_H^0} \right)^{1/3} &= \left(\frac{3}{4\pi} \frac{\zeta}{\bar{n}_H^0} \frac{\Omega_b}{\Omega_m} \frac{M}{m_p} \right)^{1/3} \\ &= 680\,{\rm kpc} \left(\frac{\zeta}{40} \frac{M}{10^8 M_\odot} \right)^{1/3} . \end{aligned} \tag{8.3}$$

Here we have taken Population II stars with $f_{\rm esc} = 8\%$ and $f_\star = 10\%$ for a fiducial estimate.

We may make a similar estimate for the size of the H II region around a quasar. For the typical quasar spectrum, $\sim 10^4$ ionizing photons are produced per baryon incorporated into the black hole, assuming a radiative efficiency of $\sim 6\%$. The overall efficiency of incorporating baryons into the central black hole is low ($< 0.01\%$ in the local Universe), but $f_{\rm esc}$ is likely to be close to unity for powerful quasars which ionize their host galaxy. Thus, quasars typically have $\zeta \sim 100$.

However, the elevated density of the IGM at high redshift implies that recombinations cannot be ignored, so this simplest method must be improved. Just before World War II, the Danish astronomer Bengt Strömgren analyzed the same problem

for hot stars embedded in the interstellar medium.[28] In the case of a steady ionizing source (and neglecting the cosmological expansion), he found that a steady-state volume (now termed a 'Strömgren Sphere') would be reached, through which recombinations are balancing ionizations:

$$\alpha_B \bar{n}_H^2 V_p = \frac{d\,N_\gamma}{dt}\ , \tag{8.4}$$

where the recombination rate depends on the square of the density and on the recombination coefficient; here we use the case-B value on the assumption that ionizing photons resulting from recombinations to the ground state would contribute to the growth of the Strömgren sphere itself (see §4.3).

To model the detailed evolution of an expanding H II region, including a non-steady ionizing source, recombinations, and cosmological expansion, we write[29]

$$\bar{n}_H \left(\frac{dV_p}{dt} - 3HV_p \right) = \frac{d\,N_\gamma}{dt} - \alpha_B \left\langle n_H^2 \right\rangle V_p\ . \tag{8.5}$$

In this equation, the mean density $\bar{n}_H \propto a^{-3}(t)$ and the angular brackets denote a volume average. Note that the recombination rate scales as the square of the density. Therefore, if the IGM is not uniform, but contains high-density clumps separated by modestly underdense voids, then the recombination time will be shorter. This is often accounted for by introducing a volume-averaged clumping factor C (which is, in general, time dependent), defined by[i]

$$C = \left\langle n_H^2 \right\rangle / \bar{n}_H^2\ . \tag{8.6}$$

Unfortunately as we will see in §8.3.1 below, the clumping factor is rather difficult to estimate robustly.

If the ionized volume is large compared to the typical scale of clumping, so that many clumps are averaged over, then equation (8.5) can be solved by specifying C. Switching to the comoving volume V, the resulting equation is

$$\frac{dV}{dt} = \frac{1}{\bar{n}_H^0}\frac{d\,N_\gamma}{dt} - \alpha_B \frac{C}{a^3}\bar{n}_H^0 V\ , \tag{8.7}$$

where $\bar{n}_H^0$ is the present number density of hydrogen. The solution for $V(t)$ around a source which turns on at $t = t_i$ is [30]

$$V(t) = \int_{t_i}^{t} \frac{1}{\bar{n}_H^0}\frac{d\,N_\gamma}{dt'}\, e^{F(t',t)} dt'\ , \tag{8.8}$$

where

$$F(t',t) = -\alpha_B \bar{n}_H^0 \int_{t'}^{t} \frac{C(t'')}{a^3(t'')}\, dt''\ . \tag{8.9}$$

We can simplify this in the high redshift limit ($z \gg 1$), where the scale factor varies as $a \propto t^{2/3}$, if we make the additional assumption of a constant C. Then, defining $f(t) = a(t)^{-3/2}$, we obtain

$$F(t',t) = -\frac{2}{3}\frac{\alpha_B \bar{n}_H^0}{\sqrt{\Omega_m} H_0}\, C\, [f(t') - f(t)] = -0.26 \left(\frac{C}{10}\right) [f(t') - f(t)]\ . \tag{8.10}$$

[i] The recombination rate depends on the number density of electrons, and in using equation (8.6) we are neglecting the small contribution made by partially or fully ionized helium.

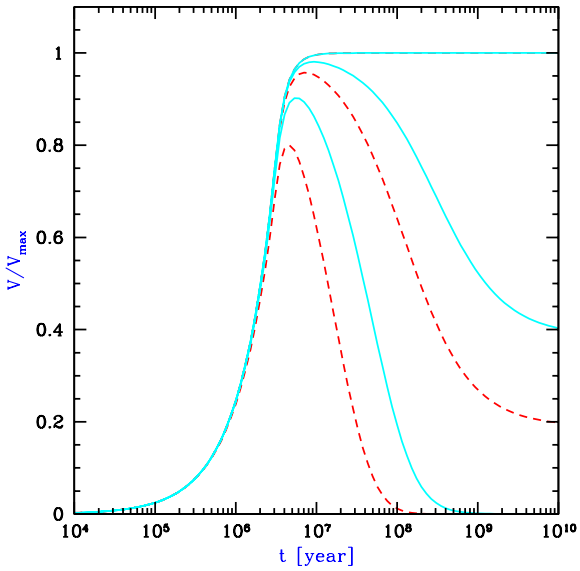

Figure 8.1 Evolution of the ionized mass for a stellar ionizing source, scaled to the maximum possible mass $M_{\rm max} = \bar{n}_H V_{\rm max} = 4\pi\bar{n}_H r_{\rm max}^3/3$ (see eq. 8.3). The solid and dashed curves assume that the sources begin shining at $z = 10$ and 15, respectively. Within each set, they take $C = 0,\ 1$, and 10, from top to bottom. The source has $\zeta = 40$ and is assumed to fade with time like $t^{-4.5}$ after a time $t_s = 3 \times 10^6$ yr, characteristic of the massive star lifetime. Figure credit: Barkana & Loeb 2001, Physics Reports, 349, 125.

One must be careful in applying equation (8.8), because the volume V is not a physical-space volume; rather, it is the comoving volume that would be filled by the ionized gas, if held at the mean density. That is, the formalism assumes that the gas inside V is completely ionized – effectively combining recombinations to the edge of the "ionized volume" – rather than allowing for the gas inside the zone to recombine uniformly. This simple model is nevertheless useful for many purposes, especially for steady sources where recombinations are relatively unimportant. We present a more rigorous model for the ionization fronts, and partial ionization inside, in §8.9.2.

Figure 8.1 shows some examples of the ionized mass for a particular model of an isolated galaxy; the results are scaled to the maximum ionized mass for the galaxy. The models take $\zeta = 40$, which makes $r_{\rm max} \sim 20 r_{\rm vir}$. They also take three possible clumping factors (from top to bottom, $C = 0,\ 1$, and 10; see §8.3.1 below) at $z = 10$ and 15 (solid and dashed curves, respectively). For this source, the ionization rate is assumed to be constant for $t_s = 3 \times 10^6$ yr, the characteristic lifetime of the massive stars that produce ionizing photons, before declining $\propto t^{-4.5}$ as these stars

die; this is a reasonable approximation to an instantaneous burst of star formation with a "normal" IMF.

Without recombinations, the ionization front reaches its maximal distance shortly after this characteristic time and remains there at later times; the result here is independent of redshift. If recombinations are allowed, the ionized mass never quite reaches its maximal value, with the shortfall increasing with redshift and clumping factor. Moreover, once they are included the ionized mass shrinks rapidly once the source dims, as recombinations destroy the ionized gas. (We remind the reader again that this does *not* mean that the front separating partially ionized and neutral gas shrinks; rather, the recombinations extend throughout the ionized volume, with that front staying more or less in place in this simple model.) Recombinations only slow down at late times as the effective recombination time exceeds the Hubble time.

One additional correction is sometimes necessary for equation (8.5): in the limit of an extremely bright source, characterized by an arbitrarily high production rate of ionizing photons, then equation 8.5 would imply that the H II region expands faster than the speed of light. At early times, the ionization front can indeed expand at nearly the speed of light, c, but only if the H II region is sufficiently small that the production rate of ionizing photons by the central source exceeds their consumption rate within the current volume. It is straightforward to take the light propagation delay into account. The general equation for the relativistic expansion of the *comoving* radius $R = (1+z)r_{\rm p}$ of an H II region in an IGM with neutral fraction $x_{\rm HI}$ is [31],

$$\frac{dR}{dt} = c(1+z)\left[\frac{\dot{N}_\gamma - \alpha_{\rm B} C x_{\rm HI} \left(\bar{n}_{\rm H}^0\right)^2 (1+z)^3 \left(\frac{4\pi}{3}R^3\right)}{\dot{N}_\gamma + 4\pi R^2 (1+z)\, c x_{\rm HI} \bar{n}_{\rm H}^0}\right], \tag{8.11}$$

where $\dot{N}_\gamma$ is the rate of ionizing photons crossing a shell of the H II region at radius R and time t (and so corresponds to the luminosity of the source at a time in the past). Indeed, for $\dot{N}_\gamma \rightarrow \infty$ the propagation speed of the proper radius of the H II region $r_p = R/(1+z)$ approaches the speed of light, $(dr_p/dt) \rightarrow c$.

8.3 GLOBAL IONIZATION HISTORY

The next level of sophistication in understanding reionization is to compute the evolution of the average neutral fraction across the entire Universe. We can obtain a first estimate for the requirements of reionization by demanding one stellar ionizing photon for each hydrogen atom in the Universe. Ignoring quasars for the time being, to zeroth order the accounting is relatively simple: the efficiency parameter ζ is simply the number of ionizing photons produced per baryon inside galaxies; thus the neutral fraction (ignoring recombinations) is

$$\bar{Q}_{\rm HII} = \zeta f_{\rm coll}, \tag{8.12}$$

where $Q_{\rm HII}$ denotes the **filling factor** of ionized bubbles (i.e., the fraction of the Universe's volume inside of H II regions) and the collapse fraction $f_{\rm coll}$ is the fraction of matter incorporated in galaxies (typically above some minimum mass

threshold determined by cooling and/or feedback). This equation assumes instantaneous production of photons, i.e., that the timescale for the formation and evolution of the massive stars in a galaxy is relatively short compared to the Hubble time at the formation redshift of the galaxy.

A simple estimate of the collapse fraction at high redshift is the mass fraction (given by equation (3.39) in the Press-Schechter model) in halos above the cooling threshold, which gives the minimum mass of halos in which gas can cool efficiently. Assuming that only atomic cooling is effective during the redshift range of reionization, the minimum mass corresponds roughly to a halo of virial temperature $T_{\rm vir} = 10^4$ K, which can be converted to a mass using equation (3.33).

The next level of sophistication is to treat each ionizing source as producing an isolated bubble and assume that their volumes add to give the total filling factor; although in fact overlap is very important, this is not a bad approximation because – neglecting internal absorption – any photons that pass into another ionized bubble propagate to its edge and help to grow it. Starting with equation (8.7), if we assume a common clumping factor C for all H II regions, then we can sum each term of the equation over all bubbles in a given large volume of the Universe and then divide by this volume. Then V can be replaced by the filling factor and N_γ by the total number of ionizing photons produced up to some time t, per unit volume. The latter quantity is simply $\zeta f_{\rm coll} \bar{n}$, which provides the emissivity of ionizing photons. Under these assumptions we convert equation (8.7), which describes individual H II regions, to an equation which statistically describes the transition from a neutral Universe to a fully ionized one:

$$\frac{dQ_{\rm H\ II}}{dt} = \zeta \frac{df_{\rm coll}}{dt} - \alpha(T) \frac{C}{a^3} \bar{n}_H^0 Q_{\rm H\ II} \ , \tag{8.13}$$

which admits the solution (in analogy with equation 8.8),

$$Q_{\rm H\ II}(t) = \int_0^t \zeta \frac{df_{\rm coll}}{dt'} \, e^{F(t',t)} dt' \ , \tag{8.14}$$

where $F(t', t)$ is determined by equation (8.10).

Although this equation appears simple, even at this level of sophistication it hides a great deal of uncertain parameters. Not only do each of the elements of ζ have large uncertainties, but they may also evolve in time; similarly, the clumping factor C depends on the pattern of ionization in the IGM. We next discuss each of these factors in turn.

8.3.1 Recombinations and the Clumping Factor

Before considering ζ, we first discuss some subtleties of the sink term in equation (8.13). First of all, the recombination coefficient is uncertain by a factor of a few through both the gas temperature (which depends on non-equilibrium processes during reionization; see §4.3.1) and an environmental factor that determines whether case-A or case-B is more appropriate. On the one hand, consider the case in which ionizations (and hence recombinations) are distributed uniformly throughout the IGM. Then case-B would be appropriate. On the other hand, in the highly-ionized low-redshift universe, most recombinations actually take place

inside dense, partially neutral LLSs because high-energy photons can penetrate inside these high-column density systems. However, the ionizing photons produced after recombinations to the ground state usually lie near the Lyman-limit (where the mean free path is small) so they are consumed inside the systems. Thus, these photons would not help ionize the IGM, and case-A would be more appropriate. Which of these regimes is more relevant depends on the details of small-scale clumping and radiative transfer.

Even more problematic is the clumping factor $C(z)$. It may seem at first that this volume averaged factor can be computed through numerical simulations. But that requires overcoming several difficult problems: *(1)* tracing the gas distribution with sufficient precision to resolve density fluctuations on the smallest scales; *(2)* correctly tracing the topology of ionized and neutral gas – because the average must be performed *only* over the ionized gas; and *(3)* correctly modeling the evolution of gas clumps during the reionization process itself.

The first problem is obvious: even leaving aside the ISM of each galaxy (which is included in $f_{\rm esc}$ in equation 8.2) the Jeans mass in the cold IGM is $< 10^5\ M_\odot$. This allows the formation of a well-defined cosmic web, as well as "minihalos," dense gas clouds that virialize but cannot cool or form stars. But, as we shall see, simulations of reionization must span ~ 100 Mpc boxes in order to adequately sample the large H II regions, requiring an enormous dynamic range. Thus, even in simulations, clumping is usually accounted for through a "subgrid" model built from semi-analytic techniques or bootstrapped from smaller simulations.

The second problem is perhaps more subtle: how do the sources and absorbers relate to each other, and how does ionization affect the small-scale clumping? For example, if low-density gas is ionized first, $C < 1$ throughout most of reionization, because all the dense gas would remain locked up in neutral, self-shielded systems (which cannot, by definition, recombine). On the other hand, on large scales the ionizing sources actually lie inside overdense regions (sheets and filaments), where the recombination rate is relatively high. The relative importance of these two features changes as reionization progresses, which makes simplified prescriptions particularly difficult to develop.

Finally, as the gas is ionized, the thermal pressure will increase and the clumps will evaporate and fade into the IGM. Studying this problem requires simulations of coupled gas dynamics and radiative transfer, which (although now possible) is difficult and highly dependent on the particular model of reionization. As an additional difficulty, the pre-reionization gas temperature is uncertain by a factor of 100 or so, making even the initial clumpiness rather uncertain as well.

Thus, while the introduction of the clumping factor is an essential approximation for many analytic models, its evaluation is rather difficult; we will describe more physically motivated approaches in §**??** below. Nevertheless, a reasonable and concrete estimate is often useful. A recent fit from simulations that ignores the second and third problems above but does resolve the proper scales is,[32]

$$C(z) = 27.466 \exp(-0.114z + 0.001328z^2). \tag{8.15}$$

8.3.2 The Ionizing Efficiency

We now move on to the source term in equation (8.13). This has two parts: $df_{\rm coll}/dt$ and the ionizing efficiency ζ. The collapse fraction for a given cosmology depends only on $M_{\rm min}$, the mass threshold for galaxy formation. The most common choice for $M_{\rm min}$ corresponds to a virial temperature $T_{\rm vir} = 10^4$ K, the threshold at which hydrogen line cooling becomes efficient for primordial gas. Above this mass, cooling and fragmentation into stars is relatively straightforward. Other choices are, however, physically plausible in certain regimes. For example, we have seen that H_2 cooling could allow Population III star formation in much smaller halos, while internal feedback within galaxies (like supernova winds) can strongly suppress star formation in galaxies near the cooling threshold, effectively raising $M_{\rm min}$.

The factor ζ is even more difficult to pin down. A star formation efficiency $f_\star \sim 10\%$ is 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 Population III stars are even more uncertain. To the extent that each halo can form only a single very massive ($\sim 10^2\ M_\odot$) star that enriches the entire halo ($> 10^6 M_\odot$), $f_\star \sim (\Omega_m/\Omega_b) M_\star/M_h < 10^{-3}$, though larger values are permissible, especially if metal dispersal is inefficient.

The UV escape fraction is small in both nearby galaxies and those at moderate, with many upper limits $f_{\rm esc} < 5\%$ and only a few positive detections. Interestingly, $f_{\rm esc}$ shows large variance between galaxies; most likely, ionizing photons are only able to escape along clear channels in the galactic ISM, which appear to be quite rare in the objects we can study. However, it could be considerably larger inside small, high-redshift galaxies, whose interstellar media can easily be shredded by radiation pressure, winds, and supernovae, clearing out large escape paths.

$N_{\rm ion}$ depends on the stellar initial mass function and metallicity. Convenient approximations are $N_{\rm ion} \approx 4000$ for $Z = 0.05 Z_\odot$ Population II stars with a present-day initial mass function, and $N_{\rm ion} < 10^5$ for very massive Pop III stars. Note, however, that the latter estimate hinges more on the high masses of these stars than on their primordial composition; metal-free stars with a normal Salpeter IMF are only ~ 1.6 times more efficient than their Pop II counterparts.

Of course, we actually expect all of these factors to evolve throughout reionization due to the feedback processes discussed elsewhere. Thus, a robust model for the filling factor $Q_{\rm H\,II}$ requires a sophisticated understanding of galaxy evolution during the cosmic dawn. This lies well beyond our powers at present, but we can make some progress by generalizing the ionizing efficiency to be a function of both time and halo mass m_h, $\zeta \equiv \zeta(m_h, t)$. The mass dependence is meant to capture internal feedback mechanisms that affect each galaxy in a deterministic fashion, like the effects of starburst winds. Of course, external feedback mechanisms – which depend on the halo's large-scale environment – require additional inputs. With this prescription, we must replace the source term in equation (8.13) with an integral over the mass function,

$$\frac{d}{dt} \int dm_h \frac{m_h}{\bar{\rho}} \zeta(m_h, t) n(m_h, t), \tag{8.16}$$

where $n(m_h, t)$ is the halo mass function.

8.4 THE PHASES OF HYDROGEN REIONIZATION

The process of the reionization of hydrogen involves several distinct stages.[33] The initial "pre-overlap" stage consists of individual ionizing sources turning on and ionizing their surroundings. The first galaxies form in the most massive halos at high redshift, which are preferentially located in the highest-density regions. Thus, the ionizing photons which escape from the galaxy itself must then make their way through the surrounding high-density regions, characterized by a high recombination rate. Once they emerge, the ionization fronts propagate more easily through the low-density voids, leaving behind pockets of neutral, high-density gas. During this period, the IGM is a two-phase medium characterized by highly ionized regions separated from neutral regions by ionization fronts. Furthermore, the ionizing intensity is very inhomogeneous even within the ionized regions.

Because these first sources are highly clustered, this early phase quickly enters the central, relatively rapid "overlap" phase of reionization when neighboring H II regions begin to overlap. Whenever two ionized bubbles are joined, each point inside their common boundary becomes exposed to ionizing photons from both sources. Therefore, the ionizing intensity inside H II regions rises rapidly, allowing those regions to expand into high-density gas which had previously recombined fast enough to remain neutral when the ionizing intensity had been low. By the end of this stage, most regions in the IGM are able to "see" many individual sources, making the ionizing intensity both larger and more homogeneous as the bubbles grow than before overlap.

During this central phase, most ionizing photons stream through the IGM without absorption, because the gas is highly-ionized. However, the proto-cosmic web makes this gas inhomogeneous, and in dense pockets of the IGM the recombination rate is much larger. These neutral regions – the high-redshift analogs of Lyman-limit systems (LLSs; see §4.4.2) – absorb any ionizing photons that strike them, preventing the H II regions from continuing to grow. Eventually, the ionized bubbles become so large that most photons strike one of these LLSs before reaching the edge of a bubble. This final "post-overlap" phase thus has slower evolution in the ionizing background, modulated by the evaporation of these LLSs, and it becomes increasingly more uniform.

Of course, this reionization process develops at different rates in different regions of the Universe; naturally, areas with an abundance of sources undergo more rapid reionization, while those with relatively few sources require input of ionizing photons from external sources. Because the galaxy population traces the underlying density field, these correspond to overdense and underdense regions, respectively. But because the galaxies are highly biased relative to the dark matter, even a modestly overdense region can undergo reionization much earlier. (In fact, if galaxies were unbiased, reionization would not occur any faster in dense regions because the increased galaxy counts would be exactly cancelled by the increased gas density!) This general march of reionization from high to low density is referred to as "inside-out" reionization (although of course, on sufficiently small scales the process is better thought of as "outside-in," since dense blobs remain partially neutral for long periods).

Figure **??** illustrates this patchiness (or "Swiss cheese topology" as it is often termed). The four panels from top left, top center, top right, and bottom left show the density of ionized hydrogen (in units of the mean) when $\bar{x}_i = 25\%,\ 50\%,\ 75\%$, and $\approx 100\%$. The bottom right panel shows the redshift $z_{\rm reion}$ at which each cell in the simulation was ionized. Note the wide distribution of ionized bubble sizes, with the largest bubbles centered around the largest clusters of galaxies in the simulation, and the tight correlation with $z_{\rm reion}$.

8.5 THE MORPHOLOGY OF REIONIZATION

Clearly, the patchiness of the ionization field – or its **morphology** – is tightly related to where galaxies had formed at high redshifts. This morphology is therefore of much interest from both theoretical and observational perspectives, and we next describe its theoretical modeling.

Given the complex physics of the sources and sinks of ionizing photons and their interaction in the IGM, it may seem that the problem must be tackled with detailed numerical simulations, and indeed much of the early work, beyond the pre-overlap stage, followed that approach. However, at its heart reionization is actually surprisingly straightforward: until the post-overlap stage, it simply requires us to count photons. Thus a great deal of progress can be made with simple analytic models.

Let us consider the simplest possible exercise: we count the number of ionizing photons produced by galaxies inside some specified volume of radius R and density δ_R and compare it to the number of hydrogen atoms. The region can only be ionized if the former exceeds the latter, or

$$\zeta f_{\rm coll}(z, \delta_R, R) > 1. \tag{8.17}$$

Here $f_{\rm coll}(z, \delta_R, R)$ is the collapse fraction within this region,

$$f_{\rm coll}(z, \delta_R, R) = {\rm erfc}\left[\frac{\delta_{\rm crit}(z) - \delta_R/D(z)}{\sqrt{2[\sigma^2_{\rm min} - \sigma^2(R)]}}\right], \tag{8.18}$$

where $\delta_{\rm crit}$ is the threshold for halo collapse (typically using the Press-Schechter criterion), the factor $D(z)$ linearly extrapolates the real density δ_R to the present day for comparison to the collapse threshold, and $\sigma^2_{\rm min}$ is the variance of the density field on the scale corresponding to the minimum mass for galaxy formation, $M_{\rm min}$. The proportionality constant ζ is the ionizing efficiency per baryon in stars (equation 8.2); here we have assumed that it is identical in every galaxy, though that is straightforward to modify as in equation (8.16).

There are two flaws to this approach. The first is that some fraction of the gas may recombine before the region is completely ionized, so more than one photon per atom is required. If such recombinations were uniform, we could account for them simply by replacing $\zeta \rightarrow \zeta/(1 + N_{\rm rec})$, where $N_{\rm rec}$ is the mean number of recombinations per baryon. In practice this is not a very good approximation, so we describe the effects of inhomogeneous recombinations later.

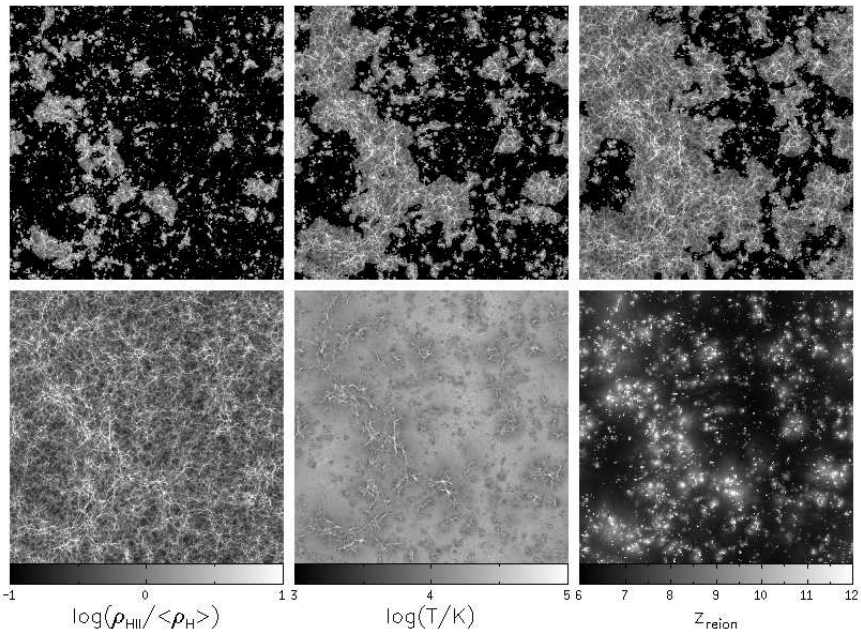

Figure 8.2 Snapshots from a numerical simulation illustrating the spatial structure of cosmic reionization in a slice of 140 comoving Mpc on a side. The simulation describes the dynamics of the dark matter and gas as well as the radiative transfer of ionizing radiation from galaxies. The first four panels (reading across from top left to bottom left) show the evolution of the ionized hydrogen density $\rho_{\rm HII}$ normalized by the mean proton density in the IGM $\langle\rho_{\rm H}\rangle = 0.76\Omega_b\bar{\rho}$ when the simulation volume is 25%, 50%, 75%, and 100% ionized, respectively. Large-scale overdense regions form large concentrations of galaxies whose ionizing photons produce joint ionized bubbles. At the same time, galaxies are rare within large-scale voids in which the IGM is mostly neutral at early times. The bottom middle panel shows the temperature at the end of reionization while the bottom right panel shows the redshift at which different gas elements are reionized. Higher-density regions tracing the large-scale structure are generally reionized earlier than lower density regions far from sources. At the end of reionization, regions that were last to get ionized and heated are still typically hotter because they have not yet had time to cool through the cosmic expansion. Figure credit: Trac, H., Cen, R., & Loeb, A. *Astrophys. J.* **689**, L81 (2009).

The second problem is the propagation of photons over large scales. Equation (8.17) is *local*, in that it only compares atoms in a region to photons generated *in the same region*. In fact, a particular patch of space may be entirely ionized by sources from outside the patch: in the extreme example, consider a spherical shell in the IGM that surrounds a galaxy. The galaxy sits inside the shell, but if the shell is sufficiently close to the galaxy it will nevertheless be ionized.

Thus, to apply equation (8.17), we require some way to adjust the scale R as needed to account for nearby sources. Fortunately, we have already studied just such a technique: the excursion set model for dark matter halos solves this very problem. In that case, the problem was that a small-scale density fluctuation might lie inside a larger-scale feature that itself may have collapsed to form a halo; in our case a small region might lie inside a larger ionized bubble. In both cases the solution is to compare the threshold (for spherical collapse or ionization) on *all* scales, working from large to small so as to include neighbors automatically, by phrasing it as a diffusion problem.

We therefore consider here the trajectory of δ_R as we move from large to small scales. We compare this smoothed density to the criterion in equation (8.17), which can be rewritten as

$$\delta_R > \delta_B(M,z) \equiv \delta_{\rm crit} - \sqrt{2}K(\zeta)[\sigma^2_{\rm min} - \sigma^2(M,z)]^{1/2}, \qquad (8.19)$$

where $K(\zeta) = {\rm erf}^{-1}(1-\zeta^{-1})$ and ${\rm erf}(x) \equiv 1 - {\rm erfc}(x)$. The barrier in equation (8.19) is well approximated by a linear function of σ^2, $\delta_B \approx B(M) = B_0 + B_1\sigma^2(M)$, where B_0 and B_1 are fitting constants. Conveniently, for this linear approximation there is an analytic solution to the diffusion problem, which we can transform into the mass function of ionized bubbles

$$\frac{dn_b}{dM} = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M^2}\left|\frac{d\ln\sigma}{d\ln M}\right|\frac{B_0}{\sigma(M)}\exp\left[-\frac{B^2(M)}{2\sigma^2(M)}\right]. \qquad (8.20)$$

This function $(dn_b/dM)dM$ provides the comoving number density of ionized bubbles with IGM mass in the range between M and $M + dM$.

The solid curves in Figure 8.3 show the resulting size distributions for a range of $\bar{x}_i$ at $z = 15$; the ordinate is the fraction of the ionized volume filled by bubbles of a given size. The most important result of these models is that bubbles grow large during the middle stages of reionization, with characteristic sizes $R_c \sim 1$, 4, 10, and 30 comoving Mpc when $\bar{x}_i = 0.2$, 0.4, 0.6, and 0.8. Comparing this to equation (8.3), it is clear that by the midpoint of reionization a *typical* ionized bubble already contains thousands of sources – overlap is indeed extremely important in determining the morphology of ionized bubbles

A second important point is the very different shape of these mass functions compared to the halo mass function, which increases toward zero mass. The barrier of equation (8.19) increases relatively rapidly toward small M, choking off the formation of small bubbles. This imprints a characteristic size R_c on the ionized bubbles. To understand this size, note that R_c is the scale at which a "typical" density fluctuation is able to ionize itself, without the input of external sources; mathematically, it is where $\sigma(R_c) \approx B$. In the large bubble limit ($B \approx B_0$), our original ionization criterion becomes

$$\zeta\, f_{\rm coll}(\delta = B_0,\, \sigma^2 = 0) = 1. \qquad (8.21)$$

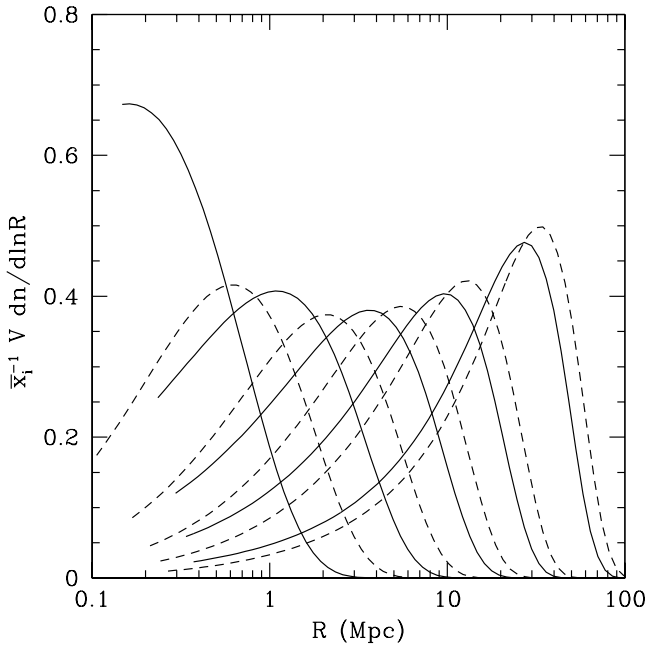

Figure 8.3 H II region size distributions at $z = 15$ in the analytic model of equation (8.20). The solid and dashed curves assume $\zeta \propto m_h^0$ and $m_h^{2/3}$, respectively. From left to right within each set, we take $\bar{x}_i = 0.05,\ 0.2,\ 0.4,\ 0.6$, and 0.8. Recombinations are assumed to be uniform throughout the IGM.

Expanding equation (8.18) to linear order, this can be written

$$\sigma(R_c) \approx B_0 \approx \frac{\bar{x}_i^{-1} - 1}{D(z) b_{\rm eff}}, \tag{8.22}$$

where $b_{\rm eff}$ is the average galaxy bias. Intuitively, a more biased galaxy population provides a larger "boost" to the underlying dark matter fluctuations, allowing larger regions to ionize themselves. The dashed curves in Figure 8.3 illustrate this effect: they show the bubble size distribution if $\zeta \propto m_h^{2/3}$, where m_h is the halo mass. This emphasizes the massive, more biased galaxies and so increases $b_{\rm eff}$. Thus, by measuring the H II region sizes, one can constrain the galaxies driving reionization.

Several properties of equation (8.20) deserve emphasis. First, at a given $\bar{x}_i$, dn_b/dM depends only weakly on redshift. This is because the shape of $f_{\rm coll}(\delta, R)$ evolves only slowly with redshift; quantitatively, $D(z)b_{\rm eff}$ is roughly constant for high-redshift galaxies, assuming that $M_{\rm min}$ is determined by a virial temperature threshold. Second, the width of $n_b(m)$ is ultimately determined by the shape of the underlying matter power spectrum, which steepens toward larger radii with a shape that is only weakly dependent on astrophysical uncertainties.

Thus, at least in this simple model, the bubble sizes depend essentially on only two parameters: the overall filling fraction of the ionized gas, $Q_{\rm HII}$, and the average

bias of the ionizing sources, $b_{\rm eff}$. Varying the overall efficiency of reionization (and hence its timing) has only a small effect on the morphology of reionization. This robustness makes the morphology an extremely useful tool in understanding the reionization process.

Finally, the similarity to the Press-Schechter halo mass function also means that most of the machinery used for halo mass functions, clustering, etc. can be carried over to describe these ionized bubbles. For example, the linear bias of H II regions, defined so that $n_b(m|\delta) = n_b(m)\,[1 + b_{\rm HII}(m)\,\delta]$ in a large region of mean overdensity δ, is[ii]

$$b_{\rm HII}(m) = 1 + \frac{B(m)/\sigma^2(m) - 1/B_0}{D(z)}. \tag{8.23}$$

Note that in this model each bubble must correspond to a region with above average density (although it can of course contain smaller underdense voids). This is obvious from equation (8.17): once the average $f_{\rm coll}(\delta = 0) = 1/\zeta$, the entire Universe must already be ionized.

However, the bias $b_{\rm HII}$ can become negative for sufficiently small bubbles. Physically, this occurs because overdense regions are farther along in the reionization process, so most small bubbles have already merged with larger H II regions. During the late stages of reionization, only the deepest voids contain galaxies isolated enough to create small bubbles. Nevertheless, the average bias of ionized gas,

$$\bar{b}_{\rm HII} \equiv Q_{\rm HII} \int dm\, \frac{dn_b}{dm} V(m) b_{\rm HII}(m) \tag{8.24}$$

is quite large throughout the early stages of reionization, attaining values $\sim$ 3–10.

As another example, each bubble must have its density equal to the barrier value at the appropriate mass (or volume). One can then generate density trajectories with the initial conditions fixed at these values and apply the usual spherical (or ellipsoidal) collapse criterion to generate the halo mass functions within each bubble; thus one can predict the galaxy populations that ionize each region of space. We explore this possibility farther in §10.6.1.

Finally, we end this section by noting that the *observed* distribution of bubble sizes differs from this "intrinsic" one. The theoretical distribution is evaluated at a single instant in cosmic time; however, real observations observe different times because of the finite speed of light. This "light-cone effect" imposes a *maximum* observable bubble size at the end of reionization, which can be estimated via similar arguments to those we have used here. Let us take the slightly simpler case of including only those photons generated within a given region of comoving radius R. Then the ionization state of that region depends only on the collapse fraction inside it. Again, reionization should be completed when this exceeds a certain critical value, corresponding to a threshold number of ionizing photons emitted per

[ii] There is one subtlety in this calculation compared to the usual halo bias. With the linear barrier fit to equation (8.19), the fractional bubble overdensity has a term $B_1\sigma_R^2/B_0$, where σ_R^2 is the mass variance on the large scale on which the bias estimate is made. This term does not scale with the dark matter density and so it spoils a linear bias estimate. Fortunately, it is large only if σ_R^2 is large (i.e., on small scales) or very close to the end of reionization, when $B_1 \gg B_0$.

baryon. There is an offset δz between the redshift at which a region of mean over-density $\bar{\delta}_{\rm R}$ achieves this critical collapsed fraction, and the redshift $\bar{z}$ at which the Universe achieves the same collapsed fraction on average.

This offset may be computed by expanding the expression for the collapsed fraction $f_{\rm coll}$ assuming small deviations (an excellent approximations on the large scales and early times relevant here), giving

$$\frac{\delta z}{(1+\bar{z})} = \frac{\bar{\delta}_{\rm R}}{\delta_{\rm crit}(\bar{z})} - \left[1 - \sqrt{1 - \frac{\sigma_{\rm R}^2}{\sigma_{\rm R_{min}}^2}}\right], \tag{8.25}$$

where again the minimum galaxy mass is $M_{\rm min}$, respectively. Obviously the offset in the ionization redshift of a region depends on its linear over-density $\bar{\delta}_{\rm R}$. Note also that equation (8.25) is independent of the critical value of the collapsed fraction required for reionization: the only redshift dependence is in $M_{\rm min}$ and is rather mild. Therefore, as with the bubble size distribution, the ionization redshift relative to its average value is nearly independent of the *timing* of reionization.

Because the density distribution narrows as R increases, the typical deviation δz decreases with R. On the other hand, the light-crossing time *increases* with R. Thus there is a critical size above which photons from the far edge of a bubble reach the observer only after the near edge of the bubble has been fully ionized. This then determines the maximum *observable* size. With the presently favored cosmological parameters, this yields ≈ 10 comoving Mpc, nearly independent of the time at which redshift occurred.

8.6 RECOMBINATIONS INSIDE IONIZED REGIONS

Incorporating inhomogeneous recombinations into the excursion set model for ionized bubbles is relatively straightforward. Each H II region obviously contains density fluctuations. Because the recombination rate increases like $(1+\delta_{\rm nl})^2$, where $\delta_{\rm nl}$ is the fully nonlinear fractional overdensity, dense clumps will remain neutral – and optically thick – longer than voids will.

We begin with the simple ansatz that there exists a threshold density δ_i below which gas is ionized and above which it is neutral.[iii] Any ionizing photons striking these dense blobs will be lost to recombinations in the neutral gas and hence are useless for increasing the filling factor of the ionized bubbles. In other words, for an H II region to continue growing, the average separation of these dense blobs must exceed the radius of the bubble. Given a model for the volume-averaged IGM density distribution, $P_V(\delta_{\rm nl})$, we can estimate δ_i by requiring the mean free path between such regions to equal the bubble radius. Clearly this threshold must increase as the bubbles grow – so that denser and denser gas is ionized.

However, ionizing more deeply into the dense gas will also increase the recom-

[iii] Of course, this cannot be exactly true, because galaxies are embedded in dense filaments, so ionizing photons do not immediately reach the voids. This model implicitly assumes that these "local" recombinations are incorporated into the escape fraction $f_{\rm esc}$.

bination rate per proton, which is

$$A_{\rm rec} = \alpha(T)\bar{n}_e(1+\delta)\int_{-1}^{\delta_i} d\delta_{\rm nl}\, P_V(\delta_{\rm nl})\,(1+\delta_{\rm nl})^2 \equiv \alpha(T)\bar{n}_e C(\delta, R_b), \tag{8.26}$$

where $C(\delta, R)$ is the *local* clumping factor within a bubble of radius R_b.[iv] The bubble can only grow if ionizing photons are produced more rapidly than recombinations consume them, or in other words if

$$\zeta\,\frac{df_{\rm coll}(z,\delta_R,\,R)}{dt} > \alpha(T)\bar{n}_e\,C(\delta,\,R), \tag{8.27}$$

The crucial point is that C depends on both the mean density of the bubble (recall that bubbles correspond to large-scale overdensities) and on its size (through δ_i). Thus, as expected from §8.3.1, inhomogeneous reionization affects the clumping factor. Moreover, the complete model is therefore both "inside-out" on large scales and "outside-in" on small scales. Recombinations become increasingly important as bubbles grow; eventually they balance ionizations and the bubble growth saturates in true cosmological Strömgren spheres.

Equation (8.27), which places a constraint on the instantaneous emissivity of ionizing photons, complements our original ionization condition, equation (8.17), which requires that the *cumulative* number of ionizing photons exceeds the total number of hydrogen atoms. In reality both conditions must be fulfilled, but in practice one of the two generally dominates. This is essentially because recombinations take over only when δ_i approaches the characteristic density of virialized objects, or in other words when LLSs dominate the mean free path, as in the lower-redshift Universe.

As a consequence, it is possible to combine the two conditions in the excursion set formalism and compute the "bubble" sizes including recombinations. However, this approach requires one conceptual shift: rather than the actual size of discrete H II regions, the radius R now corresponds to the mean free path of ionizing photons. When recombinations are unimportant, this equals the size of isolated bubbles. But once the bubbles "saturate" as Strömgren spheres, neighboring H II regions can touch – it is only that their ionizing photons will not influence each other. This is, in actuality, the same configuration that is present in the post-reionization Universe, where ionizing photons are limited by LLSs. The model therefore describes how the "bubble-dominated" topology characteristic of reionization transitions smoothly into the "web-dominated" topology of the post-reionization Lyman-α forest, albeit in an inhomogeneous manner across the Universe.

The key input parameter is obviously $P_V(\delta_{\rm nl})$, which parameterizes the IGM clumpiness (see also §4.6). In detail, the nonlinear evolution requires cosmological simulations that include coupled dark matter dynamics, gas dynamics, and radiative transfer (to account for the effects of photoheating before and during reionization). This difficult problem has not yet been solved in detail, and so approximate models

[iv] In detail, we actually require the density distribution P_V as a function of large-scale overdensity. Fortunately, in practice most large ionized bubbles (where recombinations are relevant) are very close to the mean density.

are generally used. These typically either take the post-reionization limit (where the gas is smoothed on the Jeans scale corresponding to a temperature of $\sim 10^4$ K) or appeal to a simple model for structures present before reionization (such as mini-halos).

In practice, including recombinations in this manner has a very simple effect: it imposes a *maximum size* to the "ionized regions" that corresponds to the mean free path of an ionizing photon through the inhomogeneous IGM, given the local ionizing background. Bubbles substantially smaller than this limit are almost unaffected by the LLSs, because so few of their ionizing photons strike them.

This picture has important implications for our understanding of the end of reionization. Consider, for example, the evolution of the mean specific intensity of the radiation background, $J = \epsilon\lambda/(4\pi)$, where ϵ is the emissivity and λ is the mean free path (see equation 4.40). If we ignored neutral gas inside the ionized bubbles, the mean free path would simply equal the size of the local ionized bubbble, R_b, which of course reaches infinity at the end of reionization.

Now consider how the radiation background grows at a a fixed point in the IGM, including inhomogeneous recombinations. When the point is ionized, J increases rapidly. As the sources inside the bubble ionize their surroundings – gradually adding more sources within the visible "horizon" provided by the bubble edge – J increases slowly, in proportion to R. Occasionally, however, the sources will ionize a thin wall separating a neighboring H II region. At these points, many more sources suddenly become visible and J (along with the local bubble size) increase by a large factor instantaneously. The solid curves in Figure 8.4 illustrate this series of discontinuous jumps in the ionizing background at a few different points in the IGM.

However, this series of discontinuous jumps cannot continue indefinitely: eventually, the bubble grows large enough that most ionizing photons intercept dense LLSs rather than reaching the bubble's edge. From that point, the ionizing background is regulated by the abundance of these systems rather than the global ionized fraction: in effect, the point has reached the "post-overlap" stage even if some of the IGM (at large distances from our point) remains neutral. In Figure 8.4, this is illustrated by the range of redshifts (or bubble filling factors) for which the random trajectories reach $\lambda_{\rm LLS}$.

8.6.1 The Mean Free Path at High Redshifts

Obviously, the mean free path of ionizing photons will play an extremely important role in regulating the end of reionization. Can we place any constraints on it?

This is a difficult proposition at best. Extrapolating observations at $z < 6$ (equation 4.47) imply that $\lambda \sim 7$ (1) proper Mpc at $z \sim 6$ (10); simple theoretical models predict values in this range as well. However, as the Universe becomes denser and as the ionizing background declines, the densities required to host an optically thick system approach the mean cosmic density. It is therefore not at all clear that such an extrapolation is justified.

For example, equation (4.46) tells us the density of an LLS in terms of the ionizing background. We can make a simple estimate of this background for a stellar

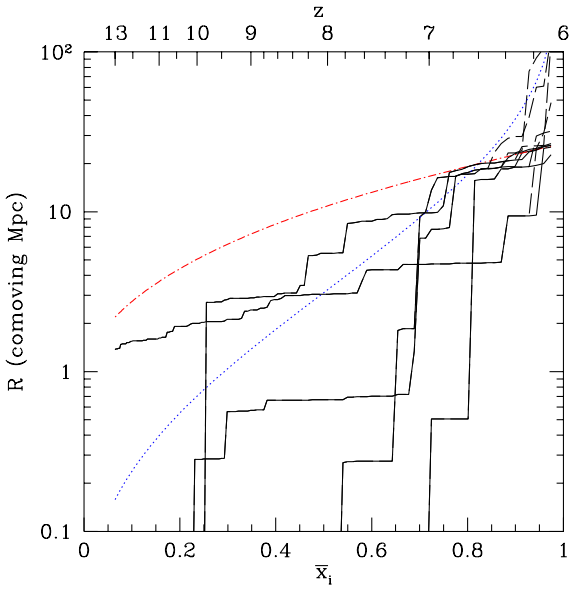

Figure 8.4 Bubble histories for several randomly generated trajectories. The vertical axis shows the bubble radius surrounding a fixed IGM point as a function of the filling factor of bubbles and z; here we arbitrarily fix ζ so that reionization completes at $z = 6$. The solid lines include the effect of inhomogeneous recombinations, while the dashed ones ignore them. These only matter when the bubbles grow larger than the mean free path of ionizing photons (near the end of reionization), so these are only distinguishable when $Q > 0.9$. The dotted and dot-dashed lines show the average bubble size R_c and $\lambda_{\rm LLS}$, respectively, in this model.

population in the context of our simple reionization model from the proper emissivity (in $\mathrm{erg\ cm^{-3}\ s^{-1}}$),

$$\epsilon \sim \zeta \frac{h\nu_{\rm HI}\rho_b}{m_p}\,\frac{df_{\rm coll}}{dt}, \tag{8.28}$$

and the mean free path λ gives,

$$\Gamma \sim \epsilon \frac{\lambda\sigma_{\rm HI}}{h\nu_{\rm HI}} \sim 2.5\times10^{-14}\,\left(\frac{\lambda}{\rm pMpc}\right)\left(\zeta\left|\frac{df_{\rm coll}}{dz}\right|\right)\ \mathrm{s^{-1}}. \tag{8.29}$$

In equation (4.46), these fiducial values imply that with $\lambda \sim 1$ proper Mpc at $z \sim 10$, $\delta_{\rm LLS} \sim 1$. Thus, the LLSs would be gas very near the mean density – presumably with much different physical properties than the dense LLSs in the moderate-redshift Universe. In fact, more detailed models that attempt to self-consistently match mean free paths of this order with IGM patches find that absorbers must lie inside weakly overdense regions.

A second concern is that the ionizing background – and hence the location of LLSs – will fluctuate across the Universe, even discounting the contrast between predominantly ionized and neutral regions. Within bubbles smaller than this mean free path, $\Gamma \propto R$ because the volume available for ionizing sources scales as R^3 while the flux from each scales as R. Thus the wide variation in bubble sizes shown in Figure 8.3 will translate into an equally wide variation in Γ, although as we have argued above the increased number of LLSs in small bubbles will not substantially affect the morphology of reionization. Moreover, even with ionized bubbles Γ has substantial (and systematic) fluctuations as they expand into low-density regions devoid of sources – although of course such regions also have fewer dense blobs capable of becoming LLSs. In practice, once the ionizing backgrounds fall low enough to be near the cosmic mean, the Γ fluctuations are more important and the optically thick systems cluster near the edges of ionized bubbles where Γ is small.

A final concern is in the uncertain amount of small-scale structure in the high-redshift IGM, which depends sensitively on the Jeans mass of this gas and hence the IGM temperature evolution. If, for example, the IGM is not significantly heated before it is ionized, the gas will be much clumpier than in the post-reionization Universe, which would render extrapolation from observations useless. We discuss these issues further in §8.10 below.

8.6.2 Maintaining Reionization

A related question (and one that existing observations can begin to answer) is whether known ionizing sources can keep the IGM ionized at a sufficiently high level. On a global scale, this requires balancing the recombination rate per unit volume with the emissivity (by number) of ionizing photons,

$$\alpha(T) C \bar{n}_H = \dot{n}_{\rm ion}. \tag{8.30}$$

Unfortunately, this equation has all of the ambiguities we have already emphasized. The choice of recombination coefficient, case-A or case-B, is not clear and depends on the nature of the absorbers as well as the underlying gas temperature (this introduces factor of two uncertainties). Moreover, the effective clumping factor C

depends on the degree to which dense regions are ionized and is somewhat degenerate with the number of ionizing photons they consume; in detail it will actually depend on the emissivity $\dot{n}_{\rm ion}$ in order to maintain net ionization equilibrium. An additional difficulty is the implicit assumption that ionizing photons are absorbed instantaneously (or equivalently that the time before absorption is much smaller than both the Hubble time and the characteristic source evolution timescale).

Nevertheless, this equation provides a simple qualitative guide to gauge whether a source population may be able to maintain the observed ionization rate in the Universe. The canonical relation for the comoving star formation density in galaxies is,

$$\dot{\rho}_\star \sim 0.003 f_{\rm esc}^{-1} \left(\frac{C}{3}\right) \left(\frac{1+z}{7}\right)^3 M_\odot \ {\rm yr}^{-1} \ {\rm Mpc}^{-3}. \tag{8.31}$$

However, converting the critical rate of ionizing photon production to a star formation rate introduces a new set of uncertainties. One substantial difficulty is the escape fraction $f_{\rm esc}$, which is uncertain to at least an order of magnitude. Others are the initial mass function (IMF) of stars, because only the most massive stars produce ionizing photons, and the metallicity, which introduces a factor of about four uncertainty in the ionizing efficiency per unit star formation rate; the relation here assumes a Salpeter IMF and solar metallicity, both of which are likely conservative and so *overestimate* the required $\dot{\rho}_\star$. Thus, without additional observational constraints on the source populations, equation (8.31) provides only a rough guide.

In order to ionize most of the IGM in the first place, the cumulative population of stars needs to produce at least one ionizing photon per hydrogen atom in the Universe. Under the same assumptions about the IMF and metallicity as used above, this condition implies a minimum comoving density of stars after reionization,

$$\rho_\star \sim 1.6 \times 10^6 f_{\rm esc}^{-1} \ M_\odot \ {\rm Mpc}^{-3}. \tag{8.32}$$

Note that this constraint does not involve the clumpiness factor, since both the number of sources and atoms scale the same way with volume.

8.7 SIMULATIONS OF REIONIZATION

So far we have discussed simplified analytic models of the reionization process. Such models ignore a large number of physical effects, including (1) the complexities of radiative transfer, such as shadowing of radiation by a dense absorber; (2) the detailed geometry of the "cosmic web" and source distribution, which is poorly approximated by spherical averaging; (3) the (possible) presence of high-energy photons that can propagate some distance through neutral gas; (4) the feedback of photoionization and photoheating on the sources of reionization and on the IGM; and (5) the nature and clustering of the dense absorbers. It is therefore necessary to develop more sophisticated numerical approaches to reionization.

8.7.1 Radiative Transfer Simulations

One option is a full cosmological simulation that attempts to include all of the relevant physics, including gravitational dynamics, hydrodynamics, and radiative

transfer. This approach is crucial for understanding many of the above issues – particularly those involving feedback of reionization itself on the gas distribution. However, it imposes daunting requirements on the simulations. Most importantly, we have seen that the relevant scales during reionization easily reach tens of Mpc, so simulating a characteristic volume requires boxes that span > 100 Mpc. On the other hand, the source halos (even discounting molecular hydrogen cooling) have masses $M \sim 10^8\ M_\odot$. Spanning both these scales – with even a single particle per galaxy – requires a dynamic range of $\sim 10^9$ (in mass), which is very difficult to achieve at present.

As a result, simulations with hydrodynamics – the most difficult of these three physics components to resolve over large dynamic ranges – typically focus on details of reionization that appear on small physical scales, such as feedback on small IGM clumps and the escape of ionizing photons from the local environment of their sources. These sorts of simulations have shown that ionization around galaxies is often highly anisotropic, due to the dense filaments along which galaxies sit, that photoheating feedback will efficiently destroy the smallest gravitationally bound clumps of baryons (or *minihalos*), and that this same feedback will moderate the clumping factor throughout the IGM. They cannot, however, describe global quantities like the average evolution of the ionized fraction or radiation background simply because the simulated volumes are too small to include more than one growing ionized bubble.

On the other hand, pure gravitational simulations of this dynamic range are relatively straightforward, and radiative transfer optimized for reionization by stellar sources (in which simply following the fate of mono-energetic ionizing photons is not a bad approximation) is relatively simple. Thus, most work to date has focused on dark matter simulations that assume a simple relation between the baryons and dark matter and apply radiative transfer to the resulting baryon field. These simulations very effectively address the the detailed geometry of the sources and cosmic web and can at least approximately address the complexities of radiative transfer and the propagation of high-energy photons, but they cannot determine how reionization feedback affects the sources or the IGM (since these are, by definition, hydrodynamic effects).

A variety of radiative transfer algorithms appear in the literature, and fortunately they seem to converge reasonably well in most circumstances. The general problem is very difficult, as computing the specific intensity $I_\nu(t,\ \mathbf{x},\ \mathbf{n},\ \nu)$ requires solving a seven-dimensional problem: time t, position $\mathbf{x}$, frequency ν, and direction of propagation $\mathbf{n}$. Furthermore, simulations can contain hundreds of thousands of sources, even excluding the diffuse light generated by IGM recombinations. Thus the complete problem is prohibitively expensive, and approximate schemes are necessary.

To simulate adequately the fact that each of the many sources illuminates its surroundings over 4π steradians requires a large number of rays. Codes typically take one of three approaches: (1) a Monte Carlo algorithm, in which a large number of photon packets are cast from the sources; (2) adaptive ray tracing, in which a small number of rays are initially cast from each source, spawning new ones as necessary to maintain the desired resolution, or (3) a field-based approach, in which photon

propagation is abstracted into a continuous field. The first is straightforward but faces the most serious convergence challenges. The second most clearly reflects the physics of the problem but is the most challenging technically. Field-based approaches are the fastest but can suffer from unusual artifacts when detailed radiative transfer effects (such as shadowing) become important.

A second question is how much to specialize the code to the particular problem of reionization. For example, the algorithm can either explicitly incorporate multi-frequency sources or focus only on counting ionizing photons. The latter is clearly significantly faster, but the former allows for non-stellar sources and is necessary to trace photo-heating accurately. Similarly, in many astrophysical contexts (including LLSs) the ionizing photons emitted during recombinations are important sources, but during reionization such photons are typically absorbed again almost immediately and so get neglected.

Still, even with this sophisticated machinery numerical simulations are ultimately limited by the same uncertainties that plague analytic models: namely, the physics *inside* high-redshift galaxies is so poorly determined that the models are descriptive but not predictive, in the sense that they can accurately predict the statistical properties of reionization given a source model but cannot from first principles generate a source model. The most sophisticated models usually take a star formation prescription calibrated to reproduce some subset of observable properties at lower redshifts. This often begins with a *Kennicutt-Schmidt law* prescription for star formation, $\dot{\rho}_\star \propto \rho_g/t_{\rm dyn} \propto \rho_g^{3/2}$, constructed to reproduce observations in which the surface density of the star formation rate scales with the gas surface density in a similar fashion, $\Sigma_{\rm SFR} \propto \Sigma_{\rm gas}^{1.4}$, over a wide range of scales in the local Universe (although recently it has become apparent that the proportionality is with the molecular gas surface density, which participates in star formation, rather than the total). This may be supplemented with a model for a multi-phase interstellar medium, feedback within the galaxy (which may drive winds into the surrounding gas), or any other physics component. A variety of such calibrations exist for lower-redshift simulations, but they have not been tested at higher redshifts.

A second problem is that these simulations cannot accurately reproduce the properties of photons sinks such as IGM clumping and LLSs, because they depend on the hydrodynamics in and around galaxies as well as feedback from photoionization. The most sophisticated models prescribe IGM clumping from higher-resolution simulations (together with some assumptions about the distribution of ionized and neutral gas and the relevant level of Jeans smoothing) and/or prescribe the distribution of LLSs based on a semi-analytic model.

The most important question is how these numerical approaches compare to the analytic models described earlier. Given all the complexities, the answer – that the analytic models fare extremely well – may be a surprise. Most importantly, the simulations show large ionized structures, with sizes comparable to those predicted, throughout most of reionization. They confirm that the filling factor of the ionized bubbles, $Q_{\rm HII}$, is by far the most important factor in determining the morphology and that the redshift is mostly unimportant. They also show that the clustering of the ionizing sources is the second most important factor and that inhomogeneous

recombinations have relatively little effect on the bubble sizes until a threshold H II region size is reached.

8.7.2 Semi-Numeric Simulations

The general agreement between these disparate approaches has inspired a set of hybrid "semi-numerical" algorithms that allow a compromise between the simplicity of the analytic models and the power of a specific realization of reionization. All of these approaches follow the same general procedure:

- First, generate the initial conditions for a cosmological simulation box (usually > 100 Mpc).
- Second, linearly evolve the density field to the desired redshift. Optionally, low-order nonlinear corrections can be applied, such as the Zel'dovich approximation.
- Third, identify the source (or dark matter halo) distribution. This is typically done by applying the excursion set approach to the specific density field of the simulation in one of two ways. One option is to use large cells and compute the expected halo abundance in each one using the analytic excursion set model. This is useful for particularly large volumes (> 1 Gpc) and/or quick and dirty estimates. A second option, useful for more detailed work and/or higher-resolution simulations, is to step through each cell in the simulation volume and smooth the density field on progressively smaller scales, identifying it as a halo whenever it crosses the spherical collapse threshold density (or an improvement upon that criterion). This mimics the random walk diffusion process used to generate the halo mass function but applies it point-by-point to account for real fluctuations in that density field. The resulting halo field does not match those of numerical simulations exactly but provides a good statistical match.
- Finally, generate the morphology of the ionized regions. Again, the density field is smoothed on progressively smaller scales around each pixel, and regions are tagged as ionized if this smoothed field exceeds the excursion set ionization criterion of equation (8.19), i.e. if the number of ionizing photons generated within the region (according to some imposed source prescription) exceeds the number of hydrogen atoms.
- Optionally, a criterion for inhomogeneous recombinations can also be included by imposing a maximum bubble size or by weighting the cells according to some estimate of the subgrid clumping and/or self-shielding.

These semi-numerical approaches thus represent a fairly direct implementation of the analytic model in specific realizations of the density field. Figure 8.5 shows that the results closely match radiative transfer simulations, at least on large scales. Clearly the broad-brush features are very similar, with ionized bubbles appearing in the same regions and growing to approximately the same sizes in each model.

Of course, the detailed shapes of the features are harder to reproduce, especially when two ionized bubbles are near to or have just overlapped with each other. In these cases, however, basing the ionization calculation on the halo field does fare significantly better.

Figure 8.6 compares four models in a more quantitative fashion through the power spectrum of the ionized fraction $P_{xx}(k)$, evaluated over the simulated volumes; this is important for many of the observables we will discuss later on. At very small scales ($k > 8h/\mathrm{Mpc}$), the models disagree, but this is largely due to shot noise in the various prescriptions. On moderate to large scales, the two radiative transfer prescriptions agree extremely well, while the semi-numeric prescriptions differ by $\sim 30\%$ late in reionization. This and other statistics show that the hybrid approaches are adequate when accuracy of this order suffices. Most importantly, the excellent agreement between this implementation of analytic reionization models and the numerical simulations suggests that – at least modulo uncertainties in the source and sink populations – existing models for the reionization process are quite robust.

The hybrid approach provide many of the advantages of large-scale simulations (especially the detailed source distribution and cosmic web topology) with computational costs orders of magnitude smaller. However, it certainly has drawbacks as well. One difficulty is that there is no a priori way to set the excursion set parameters, filtering schemes, and other details of the approach; comparison to simulations has identified the best practical schemes, but the details of the algorithms matter at the $\sim 10\%$ level. Another is that these prescriptions still invoke spherical filtering in order to paint on the ionization morphology; while the resulting configurations are certainly not themselves spherically symmetric, they do not account for complex radiative transfer effects. Third, the "photon-counting" methods we have studied so far only work for specific classes of sources in which ionizing photons are absorbed shortly after impacting neutral gas. These schemes have not yet been extended to sources with harder spectra (such as quasars, which we discuss next).

Perhaps most importantly, the semi-numeric approach cannot be used to follow the progress of reionization through time, because it does not conserve photons. Instead, the global evolution of $Q_{\mathrm{HII}}(z)$ must be prescribed externally; once that is known a series of maps can easily be generated, but they cannot then be used to infer anything about the feedback of reionization on the source population, for example. Although $Q_{\mathrm{HII}}(z)$ in radiative transfer simulations is ultimately determined by an imposed source prescriptions as well, they at least allow a self-consistent interaction of the reionization morphology with those sources.

8.8 STATISTICAL PROPERTIES OF THE IONIZATION FIELD

Figure 8.6 uses the power spectrum of the ionization fraction to compare the various simulations. The power spectrum offers a convenient way to quantify the statistical properties of a reionization model, and it can be understood intuitively based on the excursion-set model of reionization. One, relatively rigorous, approach to compute the power spectrum on a scale k is to follow two random walks, correlated on all

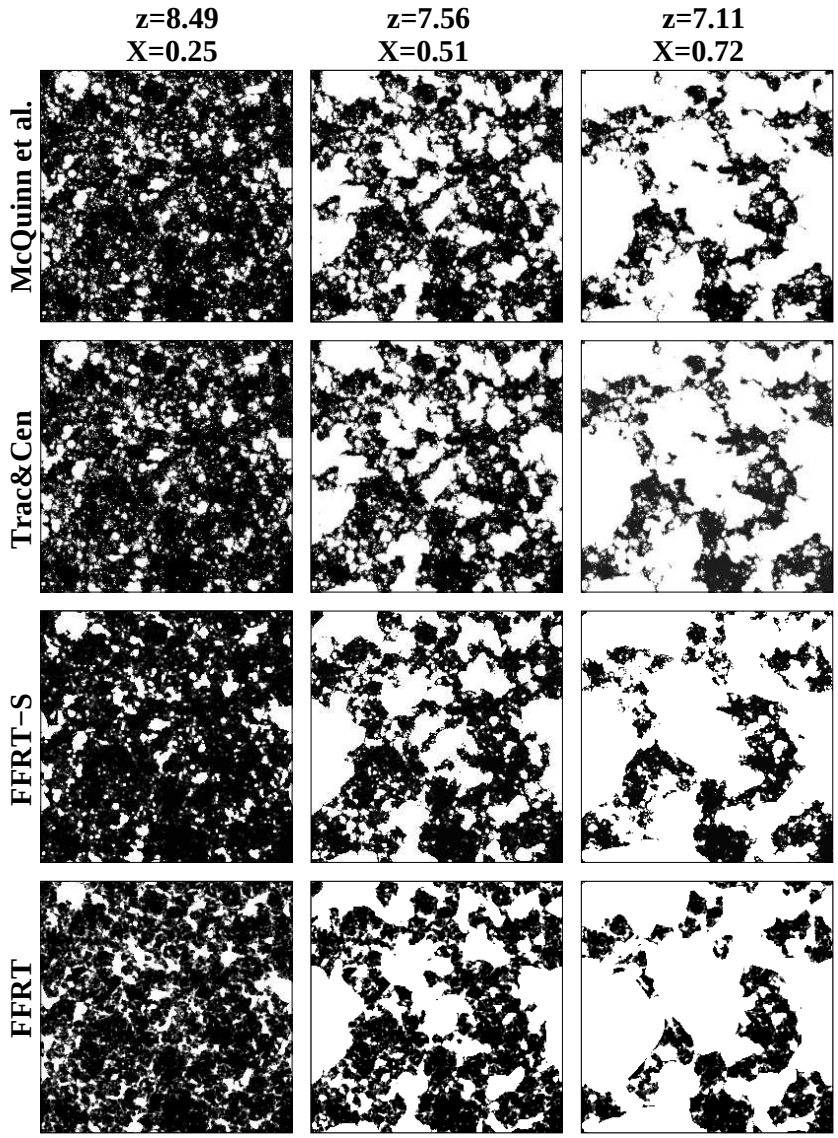

Figure 8.5 Comparison of radiative transfer and semi-numerical models of reionization. The three columns show three different times during reionization, with the filling factor of ionized bubbles (here labeled X) of 0.25, 0.51, and 0.72. The top two rows show two different radiative transfer schemes (both based on adaptive ray tracing). The bottom two rows use semi-numerical schemes: the one labeled "FFRT" uses the analytic excursions set model to predict the halo abundance, while the one labeled "FFRT-S" uses the simulated halo field itself. All four rows use exactly the same simulation volume; note the excellent agreement between the radiative transfer schemes and the close match with the semi-numerical schemes on moderate and large physical scales. The maps are 143 Mpc/h across and 0.6 Mpc deep. Figure credit: Zahn et al. (2010).

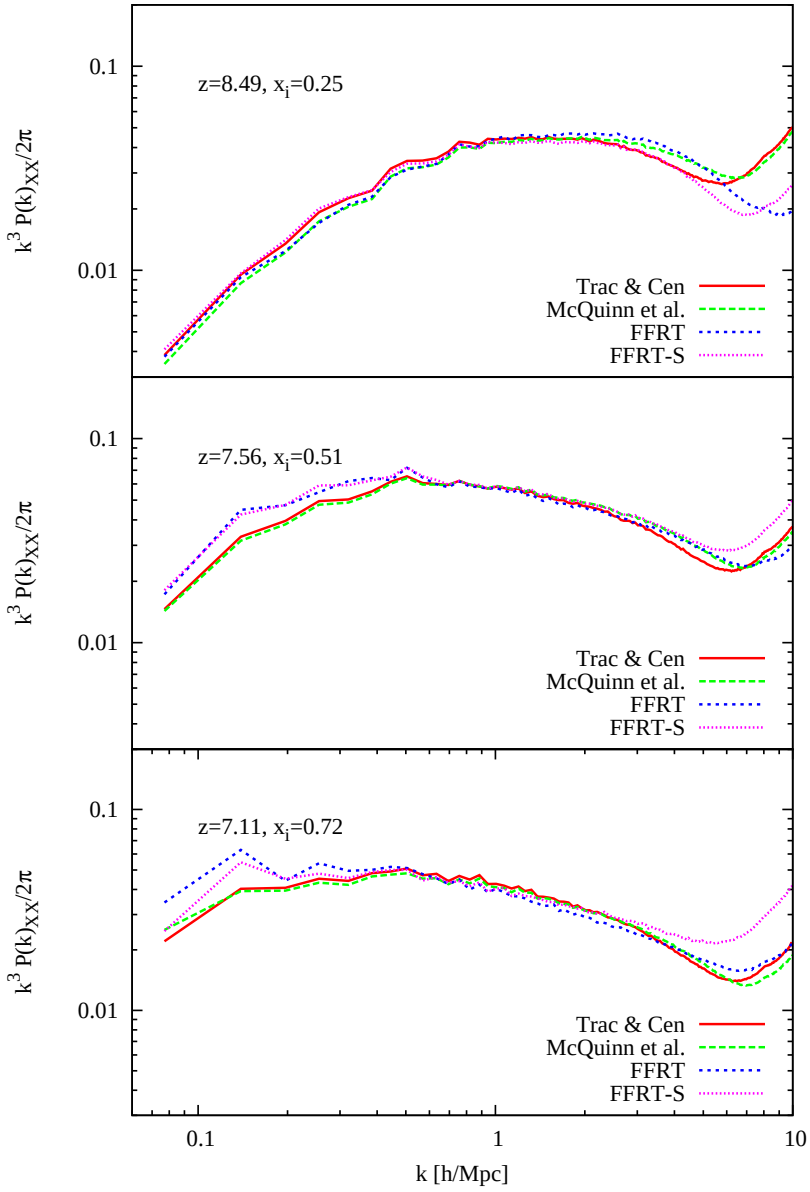

Figure 8.6 Comparison of the power spectrum of the ionization field in the radiative transfer and semi-numerical models of reionization. The three panels correspond to the columns of Fig. 8.5, and the curves correspond to the four models shown there as well. Note the close match in the predictions of all four models on scales $k < 8h/\mathrm{Mpc}$, although the semi-numerical schemes do overpredict the power on very large scales in the late stages of reionization. The differences at $k > 8h/\mathrm{Mpc}$ are due to shot noise, which differs between the schemes. Figure credit: Zahn et al. (2010).

scales $k' < k$, and determine the probability distribution of their fates inside ionized bubbles. This provides a reasonably good match to the numerical simulations.

However, we will take a simpler, approximate approach here that is informed by the simulation results. In particular, the ionization fraction is not a typical cosmological field, because it is strictly bounded to lie between zero and unity. Thus we expect the joint probability distribution of the ionized fraction at two different points to take the form

$$\langle x_i(\mathbf{r}_1)x_i(\mathbf{r}_2)\rangle = Q_{\rm HII}^2 + (1 - Q_{\rm HII})f(r/R_c), \tag{8.33}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$ and R_c the characteristic bubble size. Here f is an unknown function containing the physics of the problem, with the limits $f \to 0$ for $r \gg R_c$ and $f \to Q_{\rm HII}$ as $r \to 0$. This equation has a simple physical interpretation: if two points are separated by a distance much smaller than the size of a typical H II region they will either both be ionized by the same bubble, with probability $Q_{\rm HII}$, or both be neutral. But if $r \gg R_c$, they must reside in distinct H II regions and the probability approaches $Q_{\rm HII}^2$, with a small enhancement due to the clustering of the bubbles. The correlation function is then $\xi_{xx} = \langle x_1 x_2\rangle - Q_{\rm HII}^2$.

The second restriction arises because of the finite range of the ionized fraction: if $Q_{\rm HII} = 1$, every point must be ionized (or $x_1 = 1$ everywhere); in that case the correlations must vanish. Thus we need $\xi_{xx} = 0$ when either $Q_{\rm HII} = 0$ or 1.

The challenge lies in constructing the function f, which expresses how bubbles encompass two different points separated by a fixed distance. The correlated random walk approach implicitly computes this factor without any geometric assumptions about the bubbles. We will instead use the bubble mass function dn_b/dm, which necessitates some assumption about their structure. The simplest is of course spherical symmetry; unfortunately, this leads to an unphysical suppression in the ionized fraction near R_c. Because the excursion set formalism determines the *maximum* bubble size for which any point is a part, it does not allow for any further overlap of the bubbles. If they are all spherical, it then becomes difficult to pack them in such a way that they ionize all space – this is simplest to see in the limit in which every bubble has the same size, where reionization is then similar to packing a crate with oranges. The gaps between the oranges are impossible to remove in this situation. In reality, of course, the bubbles deform into non-spherical shapes to fill the gap, but that is difficult to model analytically.

We must therefore sacrifice rigor in order to build a simple model that approximates the final results. To do so, we split the problem into two regimes. When $Q_{\rm HII} < 0.5$, the neutral gaps are large and so reasonably well-modeled by the spherical approximation. Then, taking inspiration from the halo model, we can explicitly build the joint probability distribution by considering separately (1) the probability that a single bubble ionizes both points – the function P_1 we have already discussed, and (2) the probability P_2 that the two points are ionized by separate bubbles. In the latter, because the bubbles are relatively small at this stage, the correlations between them must be included. We then have

$$P_1(r) = \int dm\, \frac{dn_b}{dm} V_1(m, r) \tag{8.34}$$

$$P_2(r) = \int dm_1 \frac{dn_b}{dm_1} \int d^3\mathbf{r}_1 \int dm_2 \frac{dn_b}{dm_2} \int d^3\mathbf{r}_2 [1 + \xi_{bb}(r|m_1, m_2), \tag{8.35}$$

where $V_1(m,r)$ is the volume in which the center of a sphere of mass m can lie while simultaneously ionizing two points separated by r and $\xi_{bb}(r|m_1,m_2) \approx b_{\rm HII}(m_1)b_{\rm HII}(m_2)\xi(r)$ is the bubble correlation function.

Late in reionization, when $Q_{\rm HII} > 0.5$, we set $f = P_1$ in equation (8.33): while this does not include large-scale correlations, by this point the bubbles are so large that the excess correlation on scales beyond the bubble size is negligible. By doing this, we are ignoring the "two-bubble" term entirely. This means that our expression does not asymptote to a form proportional to the dark matter correlation function at late times, $\xi_{xx} \approx \bar{b}_{\rm HII}^2\xi$. However, at these late times this limit is only reached at extremely large scales, well beyond the sizes accessible to either observations or simulations. At more moderate scales, the Poisson fluctuations of the discrete bubbles dominate.

Finally, we have

$$\langle x_i x_i\rangle(r) = \begin{cases} P_1(r) + P_2(r) & Q_{\rm HII} < 0.5, \\ (1-Q_{\rm HII})P_1(r) + Q_{\rm HII}^2 & Q_{\rm HII} > 0.5, \end{cases} \tag{8.36}$$

The solid curves in Figure 8.7 compare this simple expression to the correlation function found in a semi-numeric simulation (including only the linear theory evolution in a $100h^{-1}$ Mpc box) at three different bubble filling fractions. Note the very good agreement at small and moderate scales, which suggests that this simple approach provides good intuition about the properties of the ionization field.

Also of interest is the cross-correlation between the ionized fraction and the underlying density. Again, it is relatively straightforward to construct a reasonable analytic approximation for this because the excursion set formalism is used for both the halo distribution (which via the halo model describes the density field) and the ionized bubbles. To evaluate it in detail, we can again use some simple tricks. First, suppose that the point where we evaluate the density lies inside a bubble. Then we already know the mean density of the bubble material (equal to the excursion set barrier δ_B). We can therefore approximate this part of the correlation as

$$P_{\rm in}(r) = \int dm \frac{dn_b}{dm} V_1(m,r)[1+\delta_B(m)], \tag{8.37}$$

because the x_i field is unity only inside of bubbles, where the mean density is δ_B (thus the correlation vanishes in neutral regions!).

If on the other hand the point is outside the bubble, we can approximate $\xi_{bh} \approx b(m_h)b_{\rm HII}(m)\xi$, using the linear theory expression because the distance is large. The contribution from these pairs is

$$P_{\rm out}(r) = Q_{\rm HII} - \int dm\, \frac{dn_b}{dm} V_1(m,r) + \int dm_b\, \frac{dn_b}{dm} \int d^3\mathbf{r}_b\, b_{\rm HII}\xi(r), \tag{8.38}$$

where we have used the fact that the mean halo bias is always unity to perform the integral over m_h. Here the first two terms are the fraction of space that is ionized (so that $\langle x_i\delta\rangle$ is non-zero but not contained in P_1); thus the second term in $P_{\rm out}$ cancels the first term in $P_{\rm in}$. The third term contains the correlations. As before, this term is not accurate when $Q_{\rm HII}$ is large, because $b_{\rm HII}$ encounters difficulty

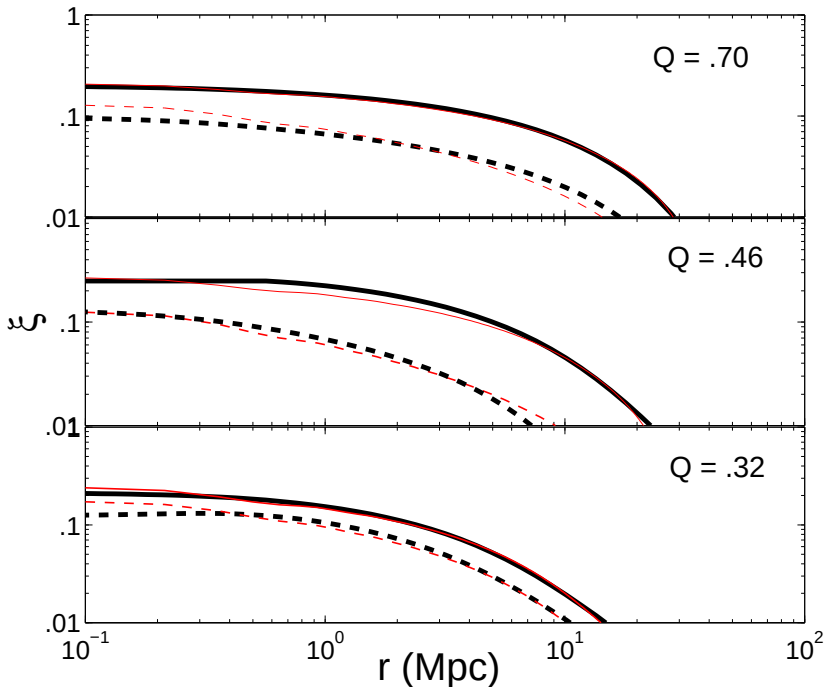

Figure 8.7 Comparison of the autocorrelation function of the ionized fraction (solid curves) and the cross-correlation function of the ionized fraction and density (dashed lines). In each case, the thick lines show our analytic approximations of eqs. (8.36) and (8.39), while the thin curves show results for a semi-numeric simulation in a $110h^{-1}$ Mpc box. The two methods are in quite good agreement at a wide range of ionized fractions. Figure credit: McQuinn, M. et al. 2005, ApJ, 630, 643.

there; however, at these times the bubble radius so large anyway that the term can be ignored. Thus we have the net approximation

$$\langle x_i \delta \rangle (\mathbf{r}_b - \mathbf{r}_h) = \begin{cases} P_{\rm in} + P_{\rm out} - Q_{\rm HII} & Q_{\rm HII} < 0.5, \\ P_{\rm in}(r) - P_1(r) & Q_{\rm HII} > 0.5. \end{cases} \quad (8.39)$$

In other words, when $Q_{\rm HII}$ is small, we must include correlations from both the bubble at $\mathbf{r}_b$ and from its neighboring bubbles (and in particular the excess correlation from their clustering). When $Q_{\rm HII}$ is large, we need only include the former effect. Subtracting the P_1 and $Q_{\rm HII}$ terms in each case isolate the excess correlations.

Figure 8.7 compares this approximate treatment of the cross-correlation with a semi-numeric calculation (thick and thin dashed curves, respectively). Again, the simple model does a rather good job over a range of ionized fractions, though it tends to underestimate the small-scale correlations because it averages over each bubble.

The important point of this simple model is that the excursion set model not only reproduces the gross properties of the bubble population but also their spatial distribution with respect to the density field. The special nature of the ionization field simplifies many of these calculations, helping to develop intuitive models that explain the simulation results. Moreover, the correlations can mostly be understood in terms of the average properties of the bubble population, because the individual H II regions are so large that nonlinear effects tend to be washed out anyway.

8.9 REIONIZATION BY QUASARS AND OTHER EXOTIC SOURCES

To this point we have focused on stellar sources of reionization, largely because galaxies seem to dominate the ionizing photon budget at $z \sim 6$. However, quasars present an interesting alternative reionization source largely because they have much harder (nonthermal) ionizing spectra than even the hottest stars. Thus, some of their photons can travel much larger distances through the IGM, and the morphology of the ionized and neutral gas will be much smoother than the sharply-defined bubbles that we have discussed.

8.9.1 How Important are Quasars to Reionization?

There are, unfortunately, very few constraints on the abundance of high-z quasars. The census of very luminous $z \sim 6$ quasars is now fairly well-determined, and their abundance seems to decline rapidly at $z > 4$. Although constraints on the shape of the luminosity function are quite weak, the total ionizing photon emissivity that comes from this population of quasars appears to fall a factor of 10-50 short of that required to maintain reionization at that time, using $C = 3$ and the arguments in §8.6.2.

Nevertheless, it is relatively easy to imagine that much smaller black holes – in particular those characteristic of the small galaxies common at high-redshifts – could play an important role in at least partially ionizing the IGM. At lower

redshifts, it is now clear that black holes are both ubiquitous and closely related to their host galaxies. The data are consistent with $M_{\rm BH} \propto \sigma^{4-5} \propto M_h^{1.33-1.67}$, where σ is the velocity dispersion. It is not clear how this relation evolves with redshift, so this scaling cannot directly be applied to the reionization era. For the purposes of a simple estimate we will simply scale $M_{\rm BH}$ to the total halo mass M_h, so that the (comoving) mass density in black holes is $\rho_{\rm BH} = f_{\rm BH} f_{\rm coll} \bar{\rho}_b$. We will scale $f_{\rm BH}$ to its local value in massive galaxies, $\sim 10^{-4}$. Because the mass function of the large dark matter halos hosting galaxies is so steep at high redshifts, taking a single value at any given redshift is reasonable, although we should recognize that it could easily either increase with redshift ($\sigma \propto (1+z)^{1/2}$) or decrease as the characteristic mass decreases.

Now let us consider how large $f_{\rm BH}$ must be in order to significantly ionize the IGM. These ionizations come from two sources: primary photoionizations from the quasar photons themselves, and secondary ionizations from the energetic secondary electrons. For a hard non-thermal spectrum $L_\nu \propto \nu^{-1}$, the latter dominate and deposit (very crudely) a fraction of the energy $f_i \sim x_{\rm HI}/3$ in ionizations. If the black holes have a radiative efficiency (relative to their rest mass) η and emit a fraction $f_{\rm UV}$ of their energy above the ionization threshold of which $f_{\rm esc,q}$ escapes the host galaxy, the expected number of ionizations per hydrogen atom is

$$N_{\rm ion} \sim 0.5 f_{\rm esc,q} \left(\frac{\eta}{0.1}\right) \left(\frac{f_{\rm UV}}{0.2}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{f_i}{1/3}\right). \tag{8.40}$$

Thus the local black hole-halo relation makes a plausible argument for a substantial contribution of quasars to reionization. Note, however, that the secondary ionizations become less and less common as $x_{\rm HI}$ decreases, so the lower-energy photons (either from quasars or stars) are still necessary to complete reionization.

Note that, unlike for stars, the escape fraction $f_{\rm esc,q}$ is likely to be quite high for quasars. Because all the quasar ionizing radiation emerges from a single source, it is much more likely to carve transparent channels in the interstellar medium of the galaxy. Moreover, much of the ionizing energy comes from relatively high-energy photons that have an easier time traversing their host galaxy without interacting.

The unresolved X-ray background offers a constraint on this scenario, because such a high-redshift quasar population would produce hard X-rays (≥ 10 keV) that free stream until today. Approximately $93 \pm 3\%$ of the SXRB has been resolved; the best estimate for the unresolved component is $J_X \sim 0.3\text{–}1 \times 10^{-12}\,{\rm erg\,s^{-1}\,cm^{-2}\,deg^{-2}}$ in the 0.5–2 keV band.

Suppose that black holes produce the high-redshift X-ray background at a median redshift z, emitting a fraction $f_{\rm HXR}$ of their energy in the $[0.5\text{–}2](1+z)$ keV range. The flux received at earth is $J = (c/4\pi)\rho_{\rm HXR}/(1+z)$, where $\rho_{\rm HXR}$ is the comoving energy density in hard X-rays produced by this early generation of black holes. Thus

$$J_{\rm X} \approx 10^{-13} f_{\rm esc,q}^{-1} \left(\frac{f_{\rm HXR}/f_{\rm UV}}{0.2}\right) \left(\frac{1/3}{f_i}\right) \left(\frac{N_{\rm ion}}{0.5}\right) \left(\frac{10}{1+z}\right) {\rm erg\,s^{-1}\,cm^{-2}\,deg^{-2}}, \tag{8.41}$$

where $f_{\rm HXR}/f_{\rm UV}$ and $\langle E \rangle$ are appropriate for a spectrum with $L_\nu \propto \nu^{-1}$ ranging from 13.6 eV to 10 keV.

Interestingly, this is just comparable to the presently-observed unresolved component. Thus, the X-ray background required if quasars *alone* reionized the Universe probably violates observed limits, but it could still make a substantial contribution to the ionization budget; thus it is certainly useful to consider scenarios in which quasars drive or affect the reionization process. Stellar mass X-ray binaries could also contribute to the X-ray production.

Moreover, it is relatively easy to imagine scenarios in which black hole accretion plays a much larger role. One possible way to evade these constraints is with a population of "mini-quasars" built from smaller black holes that may form through different channels than the very bright observable quasars. In such mini-quasars, most of the UV ionizing photons may come from an accretion disk, while hard X-rays instead come from synchrotron/inverse-Compton emission. The relative contribution of the two components is extremely uncertain, and if the non-thermal tail is relatively insignificant the X-ray background constraint would be weak.

8.9.2 Ionized Bubbles Around Quasars

The primary difference between quasars, which typically have non-thermal spectra in the UV and X-ray regimes, and stars (which are nearly thermal and so have very few high-energy photons) is that one cannot simply assume that all the ionizing photons are absorbed in a narrow region around the ionization front; instead, the higher-energy photons can propagate large distances through the intergalactic medium. The comoving mean free path of an X-ray photon with energy E is:

$$\lambda_{\rm X} \approx 11\,\bar{x}_{\rm HI}^{1/3} \left(\frac{1+z}{10}\right)^{-2} \left(\frac{E}{300\ {\rm eV}}\right)^{3}\ {\rm Mpc}; \tag{8.42}$$

thus, photons with $E > 1.5[(1+z)/15]^{1/2}\bar{x}_{\rm HI}^{1/3}$ keV propagate an entire Hubble length before interacting with the IGM. Many of the soft X-rays therefore escape the ionized bubble but deposit their energy (as ionization and heat) in the surrounding gas, "pre-ionizing" and "pre-heating" it before the ionization front itself reaches the gas.

In this case where photons leak past the "ionization front" marking the boundary between the mostly-ionized and mostly-neutral gas, the photon-counting arguments implicit to §8.2 are not sufficient. Instead we must more carefully examine the radiative transfer of ionizing photons through these regions. For simplicity, we will consider a model universe composed entirely of hydrogen; including helium complicates the equations but adds no essential new physics. As a photon travels away from its source, it encounters absorption that depends on the local ionized fraction as well as the photon energy. The total optical depth experienced by a photon with frequency ν that has traveled from a source to a radius r is

$$\tau(\nu, r, t) = \int_0^r \sigma_{\rm HI})(\nu) n_{\rm HI}(r,t) dr' \tag{8.43}$$

where $n_{\rm HI}$ is the local H I density (which may evolve either through the overall density or the neutral fraction) and where we have explicitly noted the time dependence, since the ionized region will grow as more and more photons are pumped

into it. We have also assumed that $r \ll c/H(z)$, so that we can ignore the cosmological redshift. The ionization rate at this position is then

$$\Gamma(r,t) = \int_{\nu_{\rm HI}}^{\infty} \frac{d\nu}{h\nu} \frac{L_\nu e^{-\tau(\nu,r,t)}}{4\pi r^2} \sigma_{\rm HI}(\nu) \left[1 + \left(\frac{E - E_{\rm HI}}{E_{\rm HI}}\right) f_i(E - E_{\rm HI})\right], \tag{8.44}$$

where $L_\nu = (dL/d\nu)$ is the monochromatic luminosity (per unit frequency) of the source, $E_{\rm HI} = 13.6$ eV is the ionization potential of H I, $E - E_{\rm HI}$ is the energy of the photoelectron, and $f_i(E - E_{\rm HI})$ is the fraction of this energy that goes into secondary ionization as the electron scatters through the ambient medium. This last factor describes the fate of the high-energy electrons; it is small for photons near the ionization threshold and (very roughly) approaches $f_i \sim x_{\rm HI}/3$ at high energies. A comparable fraction of the energy goes into collisional excitation of line transitions; the remainder goes into heating (see below); these fractions have been computed much more precisely using basic atomic physics.

The ionization rate at each position is then governed by

$$\frac{dn_{\rm HI}}{dt} = \Gamma n_{\rm HI} - \alpha_B(T) n_e n_{\rm HII}, \tag{8.45}$$

where $n_e = n_{\rm HII} = n_H - n_{\rm HI}$. We assume case-B recombination (i.e., local absorption of the recombination photons) for simplicity; otherwise the radiative transfer equation must include a source function for these photons as well. The "on-the-spot" approximation is usually a good one, however, because the recombination photons are emitted near the ionization threshold and so have short mean free paths. Note that we have left the clumping factor C off of equation (8.45), because integrating the ionization front evolution over space allows one to include the detailed density profile. However, it can easily be incorporated into the last term to account for clumping below the resolution of the calculation grid.

Because the recombination rate depends on temperature T (and often because the temperature is of intrinsic interest), one must also trace its evolution,

$$\frac{dT}{dt} = -2HT + \frac{2}{3}\frac{d\ln(1+\delta)}{dt} - T\frac{d\ln(2 - x_{\rm HI})}{dt} + \frac{2}{3k_B n_{\rm tot}}(Q - \Lambda), \tag{8.46}$$

where Q is the total radiative heating rate and Λ is the total radiative cooling rate. These terms describe adiabatic cooling due to the Hubble expansion, adiabatic heating or cooling due to local density inhomogeneities, the second accounts for the change in the total particle density due to ionizations and recombinations, and the fourth describes radiative cooling.

At high redshifts, radiative heating and cooling are typically dominated by photoheating and inverse Compton cooling, respectively. The former is

$$Q_{\rm ph} = \int_{\nu_{\rm HI}}^{\infty} d\nu \, \frac{L_\nu e^{-\tau(\nu,r,t)}}{4\pi r^2} \sigma_{\rm HI}(\nu) \, (E - E_{\rm HI}) \, f_h(E - E_{\rm HI}), \tag{8.47}$$

where $f_h(E - E_{\rm HI})$ is the fraction of the photoelectron energy that goes into heating. It is large for photons near the ionization threshold and (very roughly) approaches $f_h \sim 1 - 2x_{\rm HI}/3$ at high energies. The Compton cooling rate $\Lambda_{\rm comp}$ is given by,

$$\frac{2}{3}\frac{\Lambda_{\rm comp}}{k_B n_{\rm tot}} = \frac{1 - x_{\rm HI}}{2 - x_{\rm HI}} \frac{(T_{\rm CMB} - T)}{t_c}, \tag{8.48}$$

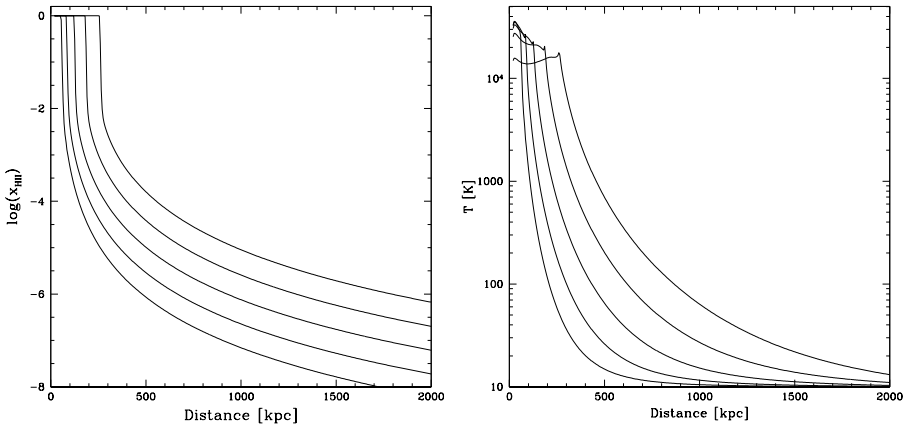

Figure 8.8 Example ionization and temperature profiles around a relatively bright quasar at $z = 10$, with $L_B = 10^9 L_\odot$. The source is assumed to emit steadily after it turns on, and the different curves take $t = 10^6$, $10^{6.5}$, 10^7, $10^{7.5}$, and 10^8 yr after ignition. The calculation assumes an initial IGM temperature of $T = 10$ K and a uniform IGM at the mean density (including helium); note that distances are measured in proper units.

where $t_c \equiv (3m_e c)/(8\sigma_T u_{\rm CMB})$ is the Compton cooling time, and σ_T is the Thomson cross section, and $u_{\rm CMB} \propto T_{\rm CMB}^4$ is the CMB energy density. The first factor on the right hand side accounts for energy sharing by all free particles.

Figure 8.8 show some example ionization and temperature profiles around a relatively bright quasar at $z = 10$ with $L_B = 10^9 L_\odot$. The source is assumed to emit steadily after it turns on, and the different curves take $t = 10^6$, $10^{6.5}$, 10^7, $10^{7.5}$, and 10^8 yr after ignition. The calculation assumes an initial IGM temperature of $T = 10$ K and a uniform IGM at the mean density. As expected, the ionization front sweeps outward over time. Behind it, the gas lies in ionization equilibrium, with $x_{\rm HI} \propto r^2$. The ionization front itself – which we will define to be the distance between which $0.1 < x_{\rm HI} < 0.9$ is narrow, but residual ionization (and heating, which can be substantial) extends several comoving Mpc from the front itself. The gas here is *not* in ionization equilibrium, as the ionization front will continue to sweep outward if the source remains luminous, and the gas outside will steadily increase in both temperature and ionized fraction.

In particular, because the recombination time in this outer region is so long (at least while the ionized fraction itself is small), the relatively low level of heating and ionization contributed by each quasar is cumulative. After many generations of AGN, the gas that remains outside of H II regions gradually becomes more and more ionized, potentially until the ionized fraction saturates at ~ 0.5 when secondary ionizations become inefficient.

This gradual ionization and heating of the otherwise untouched gas provides one

of the key difference between stellar and quasar reionization. Others are primarily driven by differences in the source luminosities and abundances: to the extent that quasars are rarer and more luminous than star-forming galaxies, they will produce larger, rarer ionized bubbles in the IGM, in which the ionized fraction and density field are less correlated. We will discuss some of the observational signatures of these differences in the later chapters.

8.9.3 Helium Reionization

So far we have focused purely on the reionization of intergalactic hydrogen. The first ionization potential of helium, 24.4 eV, is sufficiently close to that of hydrogen that helium is almost definitely singly ionized at the same time as hydrogen. However, stripping the second electron requires 54.4 eV, which is well beyond the blackbody peak of typical hot stars (although very massive metal-free stars can at least partially ionize helium). We therefore expect a significantly different ionization history for He II.

Nevertheless, many of the same tools we have already developed can be used to follow the creation of He III. Helium can easily be incorporated into the formalism of §8.9.2 by adding a multi-species network that traces the evolution of He II and He III. In practice, most high-energy photons are absorbed by He II, but (because helium is relatively rare) the secondary electron still deposits most of its energy as heat or in ionizing and exciting H I. The effects on the ionization and heating profiles (as in Figure 8.8, which does include helium) are modest and do not qualitatively affect the results.

Similar calculations for stellar sources show that only very massive metal-free stars can produce He III, although in optimistic models the He III fraction rarely rises to unity. Moreover, once these stars fade away, the He III rapidly recombines into He II because its recombination time is much shorter than that for H II (see §4.4.4) and therefore the age of the Universe t_H,

$$\frac{t^B_{\rm rec,\,He}}{t_H} \approx 0.2 \left(\frac{8}{1+z}\right)^{3/2}. \tag{8.49}$$

Thus, there may be a brief phase of ionized helium during the cosmic dawn, but it likely ends with the death of these stars.

However, radiation from quasars could provide a more sustained source of high-energy photons. We have already seen that these sources can plausibly ionize hydrogen; can they do the same for He II? The primary difference from our earlier calculation is that fast secondary electrons produced in the ionization process do not efficiently ionize He II, because its collisional ionization cross section is > 10 times smaller than that of H I (and when hydrogen is fully ionized the energy loss rate to other electrons is also much more rapid). Without secondary ionizations, the crucial parameter is the mean photon energy per ionization $\langle E_i \rangle$. If $L_\nu \propto \nu^{-1}$ from 54.4 eV to 2 keV (beyond which the IGM is optically thin), this energy is $\langle E_i \rangle \sim 200$ eV. Assuming that all of the high-energy photons ionize He II rather than H I (e.g., if stellar sources ionize the latter first), we find that the number of

ionizations per helium atom could be

$$N_{\rm ion,He} \sim 0.6 f_{\rm esc,q} \left(\frac{\eta}{0.1}\right) \left(\frac{f_{\rm UV,He}}{0.1}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{200\,{\rm eV}}{\langle E_i \rangle}\right). \quad (8.50)$$

Here $f_{\rm UV,He}$ is the fraction of the quasar's luminosity emitted above 54.4 eV. Of course, given the rapid recombination time these early quasars are unlikely to maintain more than a low level of He III in the IGM.

Despite this estimate, just as for H I the *observed* quasar high-redshift quasar population produces far fewer He II-ionizing photons. In fact estimates based on the measured quasar luminosity function predict that He II reionization must wait until $z \sim 3$, near the peak of the quasar era. Indeed, a number of lines of evidence indicate that the event occurs at roughly this time, though none are as yet definitive. We list these efforts here because they make an interesting comparison to the constraints on H I reionization that we discuss later:

- The mean optical depth of the He II Lyman-α forest appears to increase rapidly beyond $z \sim 2.8$. In §4.6 we argued that an apparently similar increase in the H I forest optical depth at $z \sim 6$ could not be interpreted in terms of reionization. But the case for helium is more secure: because the atomic number density of helium is smaller and its recombination rate is faster, its Gunn-Peterson optical depth is only $\tau \sim 14\Gamma^{-1}_{\rm HeII,-14}(1+z/4)^{9/2}$, where $\Gamma_{\rm HeII,-14}$ is the He II ionization rate in units of 10^{-14} s^{-1}. Thus, He II becomes transparent in the late stages of reionization; moreover, it does *not* have an opaque damping wing that can conceal highly-ionized regions. Additionally, reionization is accomplished by rare, bright sources whose illumination can create large (many Mpc) ionized bubbles even before the process completes. Together, these factors imply that the He II Lyman-α forest is a much cleaner probe of reionization than for H I.

- Moreover, the He II forest shows substantial fluctuations at $z > 2.8$, from being nearly opaque to very transparent. Such regions are difficult to arrange if the IGM is highly-ionized, because they would require a dearth of quasars over several hundred comoving Mpc, which is very unlikely. Unfortunately, the enormous optical depth of the H I forest at $z \sim 6$ masks the analogous fluctuations, and so this test is much more difficult to repeat with hydrogen.

- A number of measurements of the H I forest show a peak in the IGM temperature at $z \sim 3$. The most natural interpretation is photoheating from helium reionization (see §8.10).

- There is some evidence for a hardening in the metagalactic ionizing background at $z \sim 3$, as measured by the ratios of some metal lines. For example, C IV has an ionization potential just above that of He II, while Si IV has its potential just below that point. Once He II is ionized and the IGM becomes transparent to photons above 54.4 eV, we expect the abundance of C IV to decrease as well. Some (but not all) measurements show such an decrease. At $z \sim 6$, the analogous process at the H I edge should show an increase in higher ionization states, e.g., C IV, relative to low ionization states,

e.g. O I (see §4.5). Tentative evidence for such evolution does exist, but the scarcity of metal line systems at $z > 5$ and their likely positions inside highly-overdense systems complicates their interpretation in this case.

Clearly, He II reionization is at best an imperfect analog to hydrogen reionization, but it does allow us to test a number of the same ideas – particularly those relating to the ionizing background and its interaction with the IGM. It does have the key advantage of occurring at $z \sim 3$, where measurements of the H I Lyman-α forest offer a much clearer picture of the IGM. Helium reionization may therefore offer a testbed for understanding hydrogen reionization.

8.9.4 Exotic Reionization Scenarios

It is also possible that much more exotic processes – such as dark matter decay or annihilation, or primordial black hole evaporation – helped (or even completed) the reionization of the IGM. Any such exotic process that produces photons with $E > 13.6$ eV to which the IGM is opaque can also contribute to ionizing (and possibly heating) the IGM. For example, dark matter decay – even with a timescale many times the present age of the Universe – could in principle reionize the entire IGM, so long as $> 10^{-8}$ of the total rest energy of the dark matter particles went into ionization.

Although such models are quite speculative, they would produce very different patterns of reionization and so are interesting from a phenomenological perspective. For example, dark matter is fairly uniformly distributed at high redshifts, so decay would cause a nearly uniform ionizing background and hence a nearly uniform ionized fraction (moderated only by inhomogeneous recombinations and possible escape of the decay products from the source region). Annihilation would provide a clumpier source distribution but would still cause much smoother reionization than stars or quasars.

8.10 FEEDBACK FROM REIONIZATION: PHOTOHEATING

As described in §4.3.1 and §8.9.2, (some of) the excess energy deposited in the photo-electron is transformed to heat through scattering. This heating can be substantial: for a spectrum typical of a star-forming galaxy, $\Delta T \sim 12,500$–$30,000$ K (see §4.3.1), while for quasar sources one might have $\Delta T \sim 10^5$ K. We have seen that the IGM temperature is rather uncertain before reionization, but this photoheating almost certainly increases it by nearly an order of magnitude, which has a number of important consequences.

8.10.1 Photoheating and the IGM

If reionization were uniform, this dramatic heating would leave the IGM essentially isothermal. However, we have seen that in fact the process is driven by large-scale density fluctuations, with overdense regions (full of galaxies) reionized first

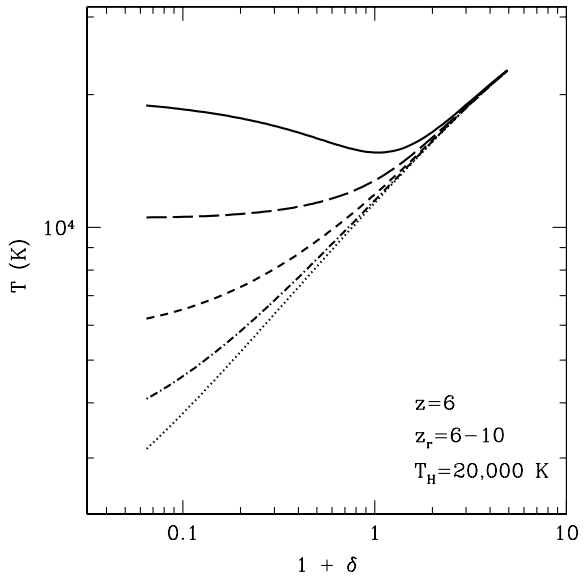

Figure 8.9 IGM temperature-density relation following H I reionization. All curves assume $z = 6$ and take a post-reionization temperature of $T_H = 20,000$ K. The solid, long-dashed, short-dashed, dot-dashed, and dotted curves set the reionization redshift at $z_r = 6$, 7, 8, 9, and 10, respectively.

and underdense regions (devoid of galaxies) reionized last (see the bottom right panel of Fig. 8.2). This translates into systematic IGM temperature fluctuations because, once reionization ends, the rapid photoheating ceases (constrained by the recombination rate within the IGM gas). Thus the overdense regions begin cooling *earlier* and have systematically cooler temperatures at later times (see the bottom center panel of FIg. 8.2).

Figure 8.9 shows this quantitatively via the IGM temperature-density relation. We show this relation at $z = 6$ computed from the excursion set reionization model of §8.5 for a variety of scenarios in which reionization ends between $z_r = 6$ and 10 (thus the different curves do not represent a sequence from one model but rather a sequence of different reionization models, with the time of observation held constant). Immediately following reionization (solid curve), the low-density voids are systematically hotter than gas near the mean density, simply because the former were the last to be ionized and so still lie near the post-reionization temperature. (Note that overdense gas is hot as well, due to the adiabatic heating from ongoing structure formation in our model.)

This kind of *inverted* temperature-density relation is strongly characteristic of "inside-out" reionization, where large-scale overdensities are ionized first, because it is characterized by the underdense voids being ionized last. Inside-out models are generic to stellar reionization, because its morphology closely traces the un-

derlying cosmic web. (Note, however, that this does not mean that *small-scale* overdensities are ionized last – in fact these LLSs typically remain neutral until the very late stages.) If, on the other hand, rare, luminous sources (such as quasars) drive reionization, the ionized bubbles correlate less strongly with the density field and the associated temperature inversion weakens (or even disappears – as appears to be the case with helium reionization at $z \sim 3$). Obviously, the IGM temperature-density relation provides a good test of the morphology of reionization.

As time passes, the expansion of the Universe causes gas at all densities to cool adiabatically. However, because underdense voids expand more rapidly than average, this cooling occurs fastest at low densities, gradually erasing the initial (inverted) temperature-density relation as time passes. Because adiabatic cooling occurs over an expansion time, the characteristic cooling timescale is the Hubble time. Thus, the interesting observational signature of inside-out reionization fades after a relatively short time, and the temperature-density relation approaches the universal asymptote in which photoheating following recombinations balances the adiabatic cooling.

Of course, reionization is also stochastic, with regions of a given density having many different reionization histories (driven by the nearby halo population). Thus, the temperature-density relation is imperfect, with scatter of $\sim 30\%$ at a given density. This scatter (and its dependence on density) also depends on the reionization model, with rarer sources inducing more scatter.

Photoheating from reionization not only increases the IGM temperature but also affects its structure: the accompanying thermal pressure increases the effective Jeans mass ($M_J \propto T^{3/2}$), evaporating existing small-scale structures and preventing accretion onto small dark matter clumps. In the diffuse IGM, this effect is usually interpreted as a decrease in the clumping factor C: when M_J is small, before reionization, very small dark matter halos can retain their baryons, causing a great deal of small-scale structure; after reionization, this gas evaporates and C decreases. Fortunately, this smoothing is relatively insensitive to the precise post-reionization temperature, because (in most models) photoheating to any reasonable temperature already increases the temperature by a very large factor.

Unfortunately, following this evolution in detail requires quite sophisticated numerical simulations that (a) resolve the small-scale IGM structure and (b) include coupled radiative transfer and hydrodynamics. To date this has only been possible in relatively small volume simulations that do not fully account for the large-scale morphology of reionization; fortunately, the insensitivity of the resulting clumping evolution to the details of the reionization process suggests that these results – in which the clumping factor can decrease by nearly a factor of two due to photoheating – are robust.

Directly observing photoheating is a challenge, especially at very high redshifts in which the Lyman-α forest is nearly saturated in absorption. In addition to the Jeans smoothing itself (which smooths out small-scale power in the forest), heating also increases thermal broadening (which smooths out the lines themselves). These manifest themselves in both statistical measures of the forest (like the power spectrum, where small-scale structure is erased) and in the lines themselves (whose shapes broaden, leaving less curvature in the spectrum). Although these techniques

have not yet been feasible at high redshifts, they are easier at moderate redshifts ($z \sim 3$) around the time of helium reionization and have been applied extensively there. Both methods have provided measurements of the evolution of the mean temperature with redshift and show heating at $z \sim 3$, of roughly the magnitude expected if quasars are responsible for the event.

However, these methods have not yet offered strong constraints on the temperature-density relation, primarily because the forest is mostly sensitive to only a narrow range of densities at any one redshift. One interesting way to avoid this problem and extend it to high redshift is by comparing constraints from multiple Lyman lines. With their weaker oscillator strengths, Lyman-β and Lyman-γ sample different parts of the density field.

8.10.2 Photoheating and Virialized Objects

Photoheating affects not only diffuse IGM gas but also gas inside of virialized objects. If such a halo has $T_{\rm vir} < 10^4$ K, photoionization will heat the gas above the escape velocity of the halo, allowing the baryons to evaporate. Moreover, once IGM gas is heated, it will ignore small dark matter potential wells, preventing the accretion of gas onto existing galaxies and suppressing subsequent star formation.

The Jeans mass in the IGM is $M_J \sim 10^5\ M_\odot$ if the gas simply cools adiabatically after decoupling from the CMB. This is far below the atomic cooling threshold ($T_{\rm vir} \sim 10^4$ K corresponds to $\sim 10^7$ Msun), so although these dark matter clumps can accrete baryons they cannot go on to form stars; instead, they will remain as dense clumps sprinkled through the IGM. Moreover, because the mass function is so steep at high redshifts, this population can contain a great deal of the collapsed mass – from $\sim 10\%$ at $z \sim 15$ to $\sim 30\%$ at $z \sim 8$. Such objects are known as **minihalos**, and their large overdensities may make them an important photon sink through the early stages of reionization.

However, these objects have shallow potential wells. As an ionization front reaches the halo, it heats the gas to $> 10^4$ K $> T_{\rm vir}$. Because the thermal pressure then exceeds the gravitational binding force, the minihalo gas escapes into the IGM through a strong evaporative wind. This escape typically occurs on roughly the sound crossing time, $\sim c_s r_{\rm vir} \sim 30(M_h/10^7\ {\rm M_{odot}})^{1/3}$ Myr at $z \sim 10$, which is much shorter than the corresponding cosmic time.

One way to parameterize the effects of the minihalos on reionization is by supplementing the clumping factor with an average $C_{\rm mh} = \langle n^2 \rangle / \langle n \rangle^2$ over the minihalo density profiles. However, the rapid time evolution during evaporation makes application of this enhanced clumping factor difficult, because one must include each minihalo for only a finite time. A simpler parameterization is with the total number of ionizing photons consumed (per minihalo atom) during the entire evaporation process. Over this time period, detailed numerical simulations show that this process typically consumes $\sim$ 3–5 ionizing photons per minihalo atom, as the high internal densities of the halos cause relatively rapid recombinations: $t^B_{\rm rec} \sim 2$ Myr for a virialized object at $z \sim 10$. Given the fraction of collapsed mass in these minihalos, this increases the number of photons per hydrogen atom required to complete reionization by about 1, potentially making minihalos as important a photon sink

as the clumped IGM itself. Fortunately, although these minihalos will be clustered and so induce inhomogeneous recombinations, numerical simulations show that treating them as approximately uniform does not introduce any significant errors.

Once a region is ionized, later formation of minihalos is strongly suppressed – even if the gas cools and recombines, because the photoionization (or indeed any other substantial heating event, such as X-rays) dramatically increases the entropy of the IGM. In this context, the quantity

$$K = \frac{T}{n^{2/3}} = 760 \left(\frac{T}{10^4\ \mathrm{K}}\right) (1+\delta)^{-2/3} \left(\frac{1+z}{10}\right)^{-2}\ \mathrm{eV\ cm^2} \qquad (8.51)$$

is usually referred to as "entropy," although the thermodynamic entropy is actually $S \propto \ln K$. Conveniently, K is conserved for any adiabatic process, including Hubble expansion flow or slow accretion; only strong shocks or radiative processes modify it. Clearly, the heating that occurs during reionization also dramatically increases the entropy; typical values, even after a substantial period of cooling and entropy release via recombination are $K_{\rm reion} > 100$ eV cm^2 at $z \sim 10$.

If this entropy is much larger than that generated by gravitational accretion onto a dark matter halo, the finite entropy "floor" will prevent gas from collapsing to high densities – essentially preventing accretion onto the halo. It is convenient to parameterize this process in terms of the entropy generated by the accretion shock at the virial radius, which provides $K_{\rm halo} \approx T_{\rm vir}/n^{2/3}(r_{\rm vir})$. Interestingly, $K_{\rm reion}/K_{\rm halo} \sim 10(T_{\rm vir}/10^4\ {\rm K})^{-1}$ for an NFW profile; thus the photoheating from reionization significantly suppresses accretion onto halos even somewhat above the usual atomic cooling threshold: numerical calculations of gas profiles (assuming hydrostatic equilibrium within the virial shock) show that only $\sim 50\%$ of the gas is able to accrete when $K_{\rm reion}/K_{\rm halo} \sim 1$, decreasing rapidly for less massive halos.

The efficiency with which photoheating suppresses accretion occurs because this process typically affects the gas while it has a small density and so efficiently imparts a large entropy to the gas. In fact, any other photoheating – even from a modest X-ray background generated by rare quasars – can substantially affect the IGM entropy, preventing the formation of minihalos even before they are ionized. We can use the estimate of equation (8.40) to examine this possibility as well: if a fraction f_h of the energy goes into heating (rather than ionization), we have

$$T_{\rm qso} \sim 20,000 f_{\rm esc,q} \left(\frac{\eta}{0.1}\right) \left(\frac{f_{\rm UV}}{0.2}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{f_h}{1/3}\right)\ \mathrm{K}, \qquad (8.52)$$

so substantial heating is clearly plausible. Even if $T_{\rm qso} \sim 1000$ K – with a very modest accompanying ionized fraction – the arguments above show that minihalo formation would be almost completely suppressed.

Figure 8.10 shows some of these effects quantitatively. The bottom panel illustrates how the entropy suppresses the collapse of gas onto dark matter halos. The uppermost solid curve shows $f_{\rm coll}$ in this model if no excess entropy is introduced, including only minihalos with $T_{\rm vir} < 10^4$ K. The dotted curves add $K = 1$ and 10 eV cm^2 (upper and lower, respectively). Even these modest levels reduce $f_{\rm coll}$ by a factor of a few to even an order of magnitude. The lower solid curve, labeled $K_{\rm IGM}(z)$ shows a *minimal* suppression due to reionization, in which the gas

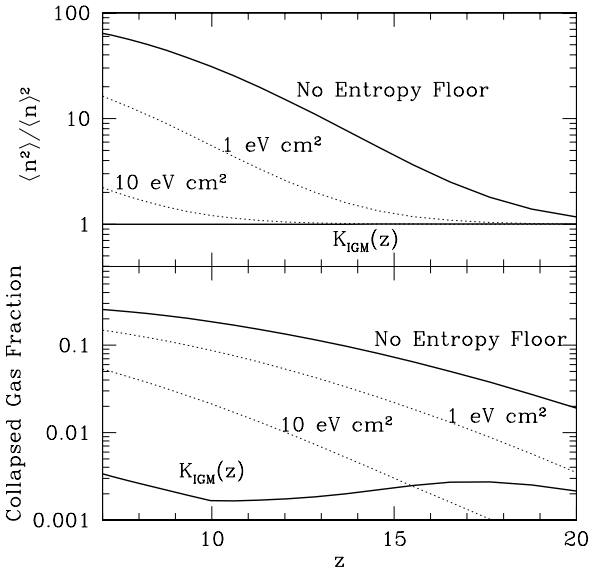

Figure 8.10 Effects of the IGM "entropy floor" on gas clumping from virialized minihalos (*top panel*) and the collapse fraction of gas onto dark matter halos (*bottom panel*. In each panel, the uppermost solid curve show the model calculation with no entropy injection. The lower solid curve, labeled $K_{\rm IGM}(z)$, shows the effect of a numerical calculation of entropy injection via photoionization and subsequent recombination (which decreases K through radiative cooling). The two dotted curves show estimates for entropy injection at fixed levels (perhaps by an X-ray background). Figure credit: Oh & Haiman 2003, MNRAS, 346, 456.)

is actually allowed to recombine for roughly a Hubble time (dramatically decreasing its entropy at high redshifts through recombination cooling). This essentially eliminates minihalo formation.

The top panel shows an estimate of the effective clumping factor, $C = \langle n^2 \rangle \, / \, \langle n \rangle^2$, when only gas inside of minihalos is included. (Thus it underestimates the *total* clumping factor that must include gas outside of virialized objects, but it more clearly shows the effect on these objects.) Again, even a relatively modest entropy injection dramatically reduces the contribution of these objects as photon sinks during reionization.

The suppression of accretion onto halos above the atomic cooling threshold is important for understanding high-redshift star formation. In detail this threshold depends on (1) self-shielding of gas within the potential well (which in turn depends upon its internal structure); (2) collisional recombination and cooling inside the halo; (3) the amplitude of the ionizing background that impinges on each halo; and (4) the relative timing of gas accretion onto the halo and the first appearance of the

ionizing background.

Fortunately, simple arguments provide an estimate for the range of halos in which accretion is eventually suppressed. Halos larger than the Jeans mass in the heated medium are essentially unaffected; this is usually parameterized by a halo circular velocity threshold, V_J (see eq. 3.32), with

$$V_{\rm J} = 81 \left(\frac{T_{\rm IGM}}{15,000\,{\rm K}} \right)^{1/2} \ {\rm km\ s^{-1}}. \tag{8.53}$$

However, the dark matter halo itself actually has an average density ~ 200 times the cosmic mean, so inside of it the gravitational force gradient is larger than in the mean density IGM. The Jeans mass evaluated with this larger density then determines the smallest halo that can accrete *any* gas, parameterized by the limiting circular velocity

$$V_{\rm lim} = 34 \left(\frac{T_{\rm IGM}}{15,000\,{\rm K}} \right)^{1/2} \ {\rm km\ s^{-1}}. \tag{8.54}$$

Halos in the range from $V_{\rm lim}$ to $V_{\rm J}$ are able to accrete some, but not their entire complement, of gas. The point at which halos are able to accrete half the expected mass is roughly the filtering mass, or time-averaged Jeans mass (see §3.2). This is somewhat *smaller* than the Jeans mass itself because the thermal pressure is lower before reionization, allowing the early phases of assembly to proceed rapidly.

Nevertheless, the effects of photoheating on high-redshift galaxies themselves are considerably weaker than these simple estimates suggest, because it takes time for the suppression to set in. Gas already close to accreting is still able to do so, because at the higher density characteristic of gas near to halos, entropy injection is less efficient. This means that photoionization feedback manifests gradually over a timescale comparable to the collapse time of dark matter halos – essentially the Hubble time. Indeed, detailed simulations show that soon after a given region is ionized, the suppression only affects halos with circular velocities $v_c < 10$km s^{-1}.

Because $V_{\rm J}$ typically lies above the atomic cooling threshold for star formation, reionization will *suppress* the formation of stars inside small galaxies. In principle this provides another test of reionization models, although as described above this suppression actually occurs gradually over a timescale comparable to the Hubble time, so it will be difficult to separate from the many other factors that affect the cosmic star formation rate. If, however, reionization is highly inhomogeneous and extended over time, the differing reionization histories in different regions of the Universe may induce variations in stellar populations whose observable effects persist to the present day. It may also have implications for understanding the wide range in stellar populations of Milky Way satellites with $V < V_J$, if some accreted gas (and formed their stars) before reionization and some after.

Chapter Nine

Galaxies at High Redshifts

The study of the first galaxies has so far been mostly theoretical, but it is soon to become an observational frontier. How the primordial cosmic gas was reionized is one of the most exciting questions in cosmology today. Most theorists associate reionization with the first generation of stars, whose ultraviolet radiation streamed into intergalactic space and broke hydrogen atoms apart in H II bubbles that grew in size and eventually overlapped. Others conjecture that accretion of gas onto low-mass black holes gave off sufficient X-ray radiation to ionize the bulk of the IGM nearly simultaneously. New observational data is required to test which of these scenarios describes reality better. The timing of reionization depends on astrophysical parameters such as the efficiency of making stars or black holes in galaxies.

Let us summarize briefly what we have learned in the previous chapters. According to the popular cosmological model of cold dark matter, dwarf galaxies started to form when the Universe was only a hundred million years old. Computer simulations indicate that the first stars to have formed out of the primordial gas left over from the Big Bang were much more massive than the Sun. Lacking heavy elements to cool the gas to lower temperatures, the warm primordial gas could have only fragmented into relatively massive clumps which condensed to make the first stars. These stars were efficient factories of ionizing radiation. Once they exhausted their nuclear fuel, some of these stars exploded as supernovae and dispersed the heavy elements cooked by nuclear reactions in their interiors into the surrounding gas. The heavy elements cooled the diffuse gas to lower temperatures and allowed it to fragment into lower-mass clumps that made the second generation of stars. The ultraviolet radiation emitted by all generations of stars eventually leaked into the intergalactic space and ionized gas far outside the boundaries of individual galaxies.

The earliest dwarf galaxies merged and made bigger galaxies as time went on. A present-day galaxy like our own Milky Way was constructed over cosmic history by the assembly of a million building blocks in the form of the first dwarf galaxies. The UV radiation from each galaxy created an ionized bubble in the cosmic gas around it. As the galaxies grew in mass, these bubbles expanded in size and eventually surrounded whole groups of galaxies. Finally, as more galaxies formed, the ionized bubbles overlapped and the initially neutral gas in between the galaxies was completely reionized.

Thus, it is galaxies – distant ancestors of our own Milky Way – that formed the building blocks of large-scale structure during the reionization era (and likely most of the cosmic dawn). In this chapter we will examine these objects in some detail, from both theoretical and observational perspectives. Although the above progression of events is plausible, at this time it is only a conjecture in the minds

of theorists that has not yet received confirmation from observational data. The exploration of the reionization epoch promises to be one of the most active frontiers in cosmology over the coming decade. We are now in a position to understand the first pillar of these efforts: direct observations of galaxy populations.

What makes the study of the first galaxies so exciting is that it is a work in progress. Scientific knowledge often advances like a burning front, in which the flame is more exciting than the ashes. It would obviously be rewarding if our current theoretical ideas are confirmed by future observations, but it might even be more exciting if these ideas are modified. In the remaining sections of this chapter, we describe the basic tools that can be used to understand the role of high-redshift galaxies during the cosmic dawn.

9.1 TELESCOPES TO OBSERVE HIGH-REDSHIFT GALAXIES

9.1.1 The Hubble Deep Field and its Follow-ups

In 1995, Bob Williams, then Director of the Space Telescope Science Institute, invited leading astronomers to advise him where to point the Hubble Space Telescope (HST) during the discretionary time he received as a Director, which amounted to a total of up to 10% of HST's observing time.[i] Each of the invited experts presented a detailed plan for using HST's time in sensible, but complex, observing programs addressing their personal research interests. After much of the day had passed, it became obvious that no consensus would be reached. "What shall we do?" asked one of the participants. Out of desperation, another participant suggested, "Why don't we point the telescope towards a fixed non-special direction and burn a hole in the sky as deep as we can go?" – just like checking how fast your new car can go. This simple compromise won the day since there was no real basis for choosing among the more specialized suggestions. As it turned out, this "hole burning" choice was one of the most influential uses of HST as it produced the deepest image we have so far of the cosmos.

The Hubble Deep Field (HDF) covered an area of 5.3 squared arcminutes and was observed over 10 days (see Figure 9.1.1). One of its pioneering findings was the discovery of large numbers of high-redshift galaxies at a time when only a small number of galaxies at $z > 1$ were known. The HDF contained many red galaxies with some reaching $z > 6$. The wealth of galaxies discovered at different stages of their evolution allowed astronomers to estimate the variation in the global rate of star formation per comoving volume over the lifetime of the universe.

Subsequent incarnations of this successful approach included the HDF-South and the Great Observatories Origins Deep Survey (GOODS). A section of GOODS, occupying a tenth of the diameter of the full moon (equivalent to 11 square arcminutes), was then observed for a total exposure time of a million seconds to create the Hubble Ultra Deep Field (HUDF), the most sensitive deep field image in visible light to date.[ii] Red galaxies have been identified in the HUDF image up to a red-

[i] Turner, E. private communication (2009).

[ii] In order for galaxy surveys to be statistically reliable, they need to cover large areas of the sky.

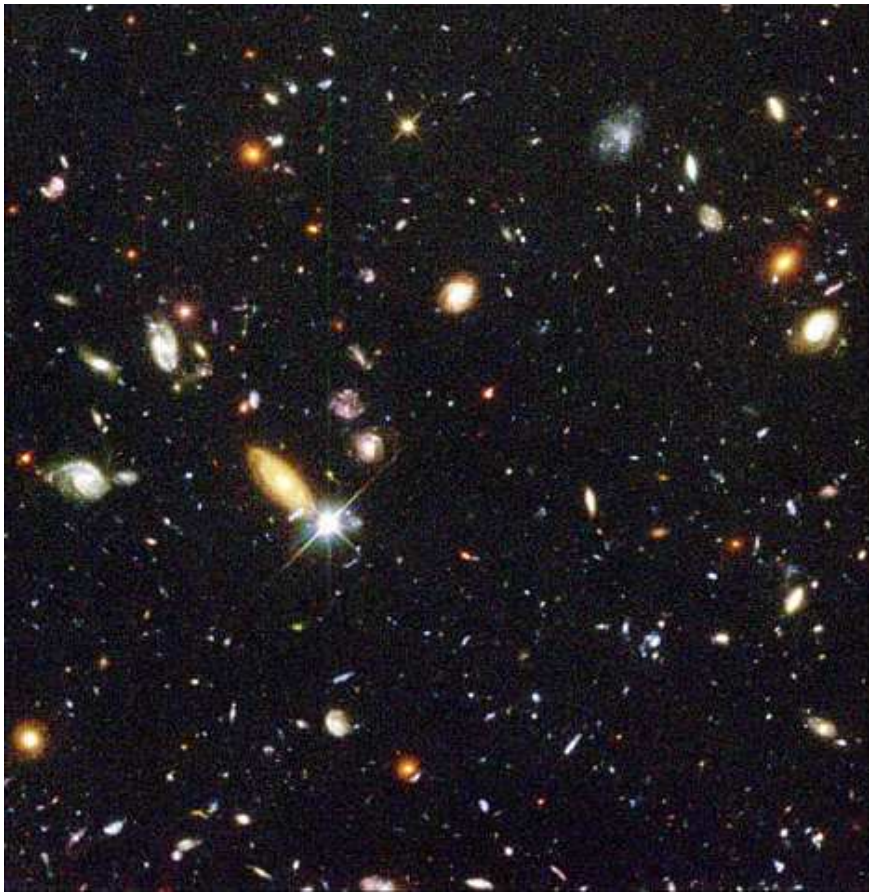

Figure 9.1 The first Hubble Deep Field (HDF) image taken in 1995. The HDF covers an area 2.5 arcminute across and contains a few thousand galaxies (with a few candidates up to a redshift $z \sim 6$). The image was taken in four broadband filters centered on wavelengths of 3000, 4500, 6060, and 8140Å, with an average exposure time of ~ 0.127 million seconds per filter.

shift of $z \sim 8$, and possibly even higher, showing that the typical UV luminosity of galaxies declines with redshift at $z > 4$. Most of the data we will discuss in this chapter ultimately comes from the HDF and HUDF.

9.1.2 Future Telescopes

The first stars emitted their radiation primarily in the UV band, but because of intergalactic absorption and their exceedingly high redshift, their detectable radiation is mostly observed in the infrared band. The successor to the Hubble Space Telescope, the James Webb Space Telescope (JWST), will include an aperture 6.5 meters in diameter, made of gold-coated beryllium and designed to operate in the infrared wavelength range of 0.6–28μm (see Figure 9.2). JWST will be positioned at the Lagrange L2 point, where any free-floating test object stays in the opposite direction to that of the Sun relative to Earth. JWST's large aperture and position outside the Earth's atmosphere makes it particularly well-suited to detect the faint, compact galaxies we expect to exist during the Cosmic Dawn and possibly discover "smoking gun" signatures of Population III stars, such as strong UV sources with no metal lines or strong He II recombination lines (see §5.4).

Several initiatives to construct large infrared telescopes on the ground are also underway. The next generation of ground-based telescopes will have effective diameters of 24-42 meters, roughly three times larger than the largest existing optical/near-infrared telescopes; examples include the European Extremely Large Telescope,[34] the Giant Magellan Telescope,[35] and the Thirty Meter Telescope,[36] which are illustrated in Figure 9.3. Along with JWST, they will be able to image and survey a large sample of early galaxies.

Additional emission at submillimeter wavelengths from molecules (such as CO), ions (such as C II), atoms (such as O I), and dust within the first galaxies would potentially be detectable with the future Atacama Large Millimeter/Submillimeter Array (ALMA).[37] This array will contain sixty six 7 to 12 meter antennas positioned at very high altitudes in Chile, in order to see past the strong atmospheric absorption at millimeter and submillimeter wavelengths. It is perfectly positioned to observe emission from dust and heavy elements in the early Universe.

Many other instruments are under development, complementing the direct views of the galaxies that one can obtain with these telescopes. For example, given that these galaxies also created ionized bubbles during reionization, their locations should be correlated with the existence of cavities in the distribution of neutral hydrogen. Within the next decade it may become feasible to explore the environmental influence of galaxies by using infrared telescopes in concert with radio observatories that will map diffuse hydrogen at the same redshifts (see §11 and 12.3).

Counts of galaxies in small fields of view suffer from a large cosmic variance owing to galaxy clustering.

Figure 9.2 A full scale model of the James Webb Space Telescope (JWST), the successor to the Hubble Space Telescope (http://www.jwst.nasa.gov/). JWST includes a primary mirror 6.5 meters in diameter, and offers instrument sensitivity across the infrared wavelength range of 0.6–28μm which will allow detection of the first galaxies. The size of the Sun shield (the large flat screen in the image) is 22 meters$\times$10 meters (72 ft$\times$29 ft). The telescope will orbit 1.5 million kilometers from Earth at the Lagrange L2 point.

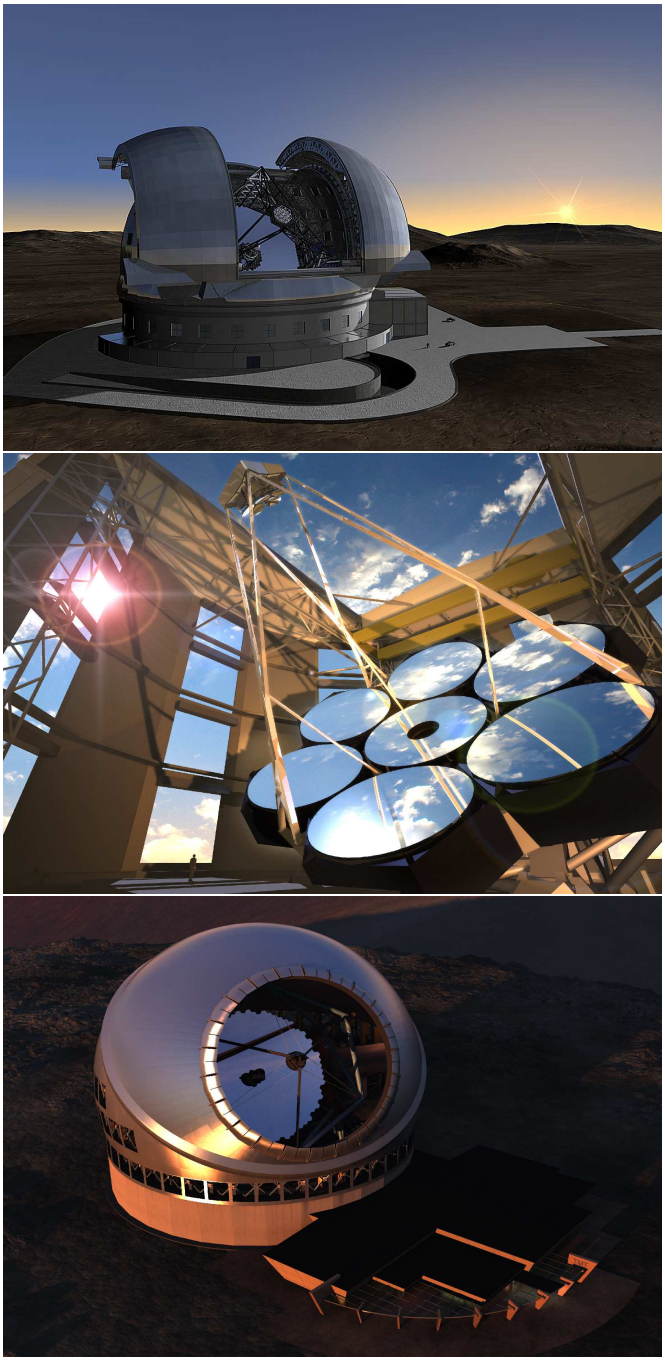

Figure 9.3 Artist's conception of the designs for three future giant telescopes that will be able to probe the first generation of galaxies from the ground: the European Extremely Large Telescope (EELT, top), the Giant Magellan Telescope (GMT, middle), and the Thirty Meter Telescope (TMT, bottom). Images credits: the European Southern Observatory (ESO), the GMT Partnership, and the TMT Observatory Corporation.

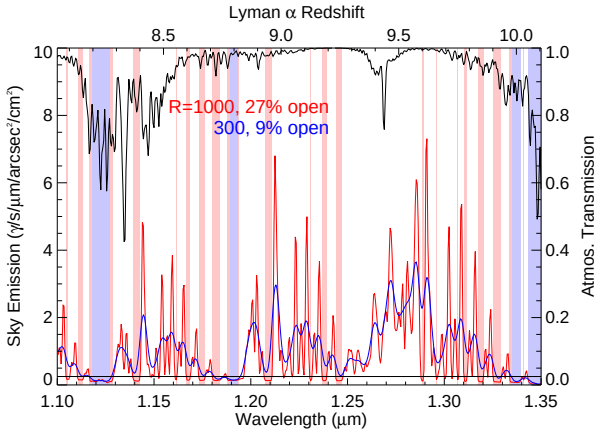

Figure 9.4 "Windows" in the J-band night-sky spectrum. The black line indicates the transmission of the night sky (scale on right). The two lines at the bottom show the night-sky spectrum at two resolutions ($R = 1000$ and 300; scale on left). The vertical shading shows regions where the emission is less than 1/3 of the mean value. Figure credit: Barton, E. J. et al., *Astrophys. J.*, **604**, L1 (2004).

9.2 METHODS FOR IDENTIFYING HIGH-REDSHIFT GALAXIES

Much of the baryonic mass in the Universe assembled into star forming galaxies after the first billion years in cosmic history. Consequently, the highest-redshift galaxies are a rarity among all faint galaxies on the sky. A method for isolating candidate high-redshift galaxies from the foreground population of feeble lower-redshift galaxies is required in order to identify targets for follow-up studies.

9.2.1 Lyman-α Emitters

One technique makes use of narrow-band imaging to identify galaxies for which highly-redshifted line emission falls within the selected band. An object that is bright in the narrowband but faint (or, for these applications, usually invislble) in nearby broadband measurements can be identified as a line emitter. This method is typically applied to the Lyman-α line, which is often very strong because most ionizing photons absorbed by the galaxy's interstellar medium (ISM) are reprocessed into Lyman-α line photons through recombinations. However, it is also highly sensitive to the gas geometry and kinematics and can be extinguished by dust. The galaxies detected by this technique are termed *Lyman-α emitters (LAEs)*. The primary challenges with this approach are:

- *The infrared night sky:* Terrestrial telescopes suffer from substantial atmospheric absorption and strong night-sky lines in the infrared bands (primarily from OH ions and water vapor). Figure 9.4 shows the night sky in the relevant spectral range, including both atmospheric absorption and night-sky

emission lines. The vertical shaded columns show "windows" where the emission lines are below 1/3 of the average. The dark and light regions take moderate and high resolution bands, respectively (with $R = \lambda/\Delta\lambda = 300$ and 1000). These open bands cover only 16% and 27% of the available spectrum, respectively, indicating that this technique can only be used in particular redshift ranges. So far, the most commonly utilized are at $z \sim 6.6,\ 7,\ 7.7,$ and 8.5.

- *Contamination from lower-z line emitters:* Galaxies have many other emission lines, of course, some of which can be very strong. Of particular concern are Hα, [O III], Hβ, and [O II] lines. Such contaminants can be ruled out by identifying other emission lines (unlikely to be possible for a true LAE, but very plausible for the lower-redshift interlopers) or by measuring the continuum emission (obviously very difficult for a faint source). If only a single line is visible, the shape can help determine whether the object is truly an LAE: as shown in Figure 9.5, observed Lyman-α lines in galaxies nearly always have asymmetric profiles, with a sharp cutoff on the blue side (due to IGM absorption) and a long tail to the red side (due to radiative transfer effects). Metal lines, on the other hand, are generally very symmetric.

- *Interpretation and followup:* Finally, although this method efficiently finds galaxies at high redshifts, it provides little direct physical information – only a single line luminosity, which as we will see in chapter 10 is heavily dependent on dust, the ISM clumpiness and dynamics, and the IGM ionization state. Even deep followup observations typically detect little or no stellar continuum emission.

To date, LAE surveys have detected many high-z sources, but their interpretation is still debated. We will return to the Lyman-α line as an important cosmological probe in chapter 10.

9.2.2 Lyman-Break Galaxies

The second observational technique adopts several broad bands to estimate the redshifts of galaxies based on the strong spectral break arising from absorption by intergalactic (or galactic) neutral hydrogen along the line-of-sight to the source. As we saw in chapter 4, the IGM is optically thick to Lyman-α photons at high redshifts. Thus, little or no flux should be detectable shortward of $1216\text{Å}(1+z)$ (irrespective of the history of reionization). For example, to identify a galaxy at $z = 6$ one needs two filters: one above and the other below the Lyman-α break at $7 \times 1216 = 8512$Å. The relevant bands are i' (centered at ~ 9000Å) and z' (centered at ~ 8000Å) of HST, as illustrated in Figure 9.6. This method was first used at lower redshifts, $z \sim 3$–4, where the intergalactic HI column density is smaller and so the related Lyman-limit break at 912Å was instead adopted to photometrically identify galaxies. The 912Å break is not observable at source redshifts $z > 6$, because it is washed out by the strong Lyman-α absorption at lower redshifts. The sources detected by this techniques are termed *Lyman-break galaxies (LBGs)*.

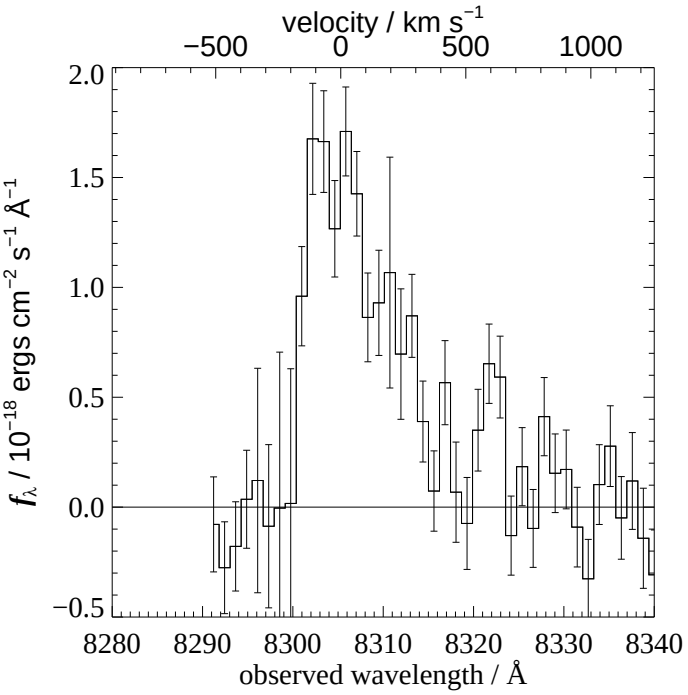

Figure 9.5 The Lyman-α emission line of a typical i'-dropout galaxy SBM03# at $z = 5.83$. Figure credit: Stanway, E., et al. *Astrophys. J.* **607**, 704 (2004).

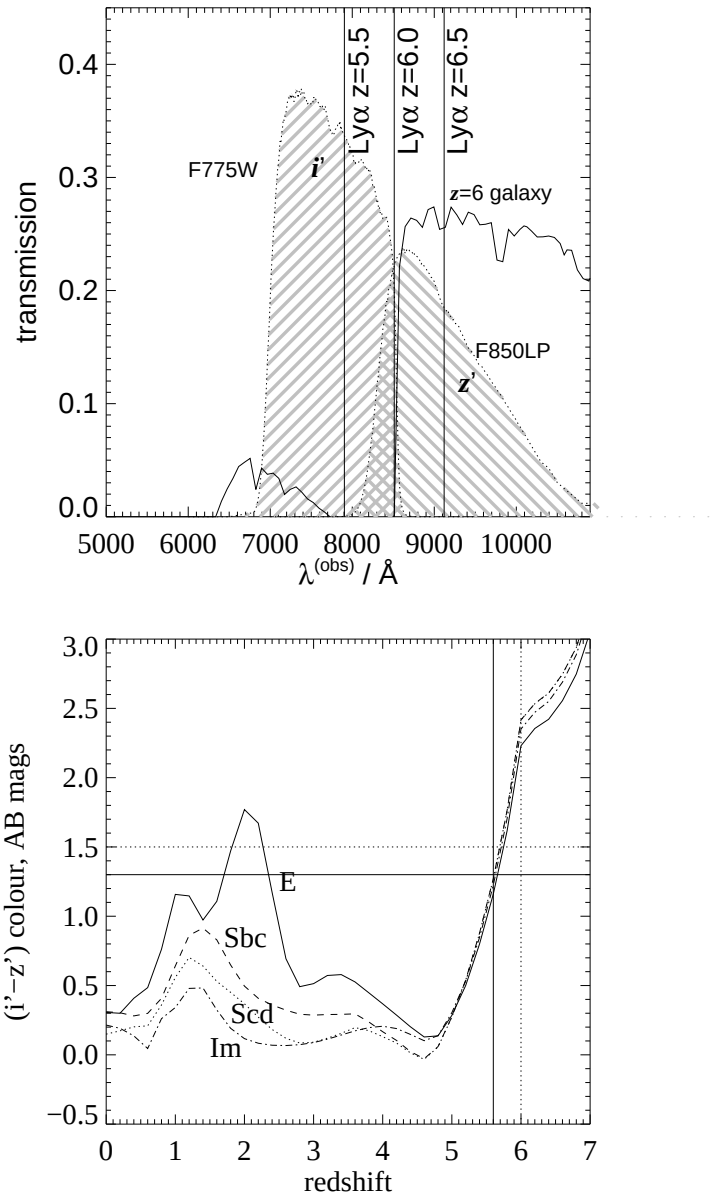

Figure 9.6 *Top panel:* The i' and z' bands of HST (shaded regions) on top of the generic spectrum from a galaxy at a redshift $z = 6$ (solid line). The Lyman-α wavelength at various redshifts is also shown. *Bottom panel:* Models of the color-redshift tracks for different types of galaxies with non-evolving stellar populations. The bump at $z \sim 1$–2 arises when the Balmer break or the 4000Å break redshift beyond the i'-filter. Synthetic models indicate that that the Balmer break takes $\sim 10^8$ years to establish, providing a measure of the galaxy age. Figure credit: Bunker, A. et al., preprint arXiv:0909.1565, (2009).

The key challenge of observers is to obtain a sufficiently high signal-to-noise ratio that LBGs can be safely identified through the detection of a single redder band. Figure 9.6 illustrates how a color cut of $(i' - z')_{\rm AB} > 2.3$ is effective at selecting sources at redshifts $z > 6$. The reliability of this dropout technique in rejecting low-redshift interlopers can only be tested through spectroscopic observations. The i'-drop spectra typically show a single emission line at the Lyman-α wavelength, with no significant continuum; as in Figure 9.5, the lines are typically asymmetric and can indicate clearly the source redshift. However, only a fraction of galaxies have Lyman-α lines and spectroscopic followup is often difficult.

The NIRSpec spectrograph on JWST covers observed wavelengths in the range $0.8 - 5\mu$m and is ideally suited for the task of identifying the redshifts of distant galaxies. This instrument will have the sensitivity to detect the rest-frame UV and optical continuum emission over the full range of emission lines from Lyman-α (1216Å) to Hα (6563Å) for galaxies at $z \sim 6$. Analogous studies of galaxies at $z \sim 3$ with HST have produced a detailed understanding of the internal properties of these galaxies.

9.2.3 Finding Faint Galaxies With the First Gamma-Ray Bursts

Traditional methods of finding galaxies, including both the LAE and LBG technique, select galaxies above a given luminosity threshold and so are biased toward identifying the brightest galaxies. However, as we will see below much of the activity at high redshifts likely occurs in faint galaxies far below the luminosity threshold of even extremely deep observations like the HUDF. Is there any way to find more typical galaxies?

Remarkabely, the best way is to find individual sources rather than the collective emission of galaxies. Explosions of individual massive stars (such as supernovae) can outshine their host galaxies for brief periods of time. The brightest among these explosions are *Gamma-Ray Bursts (GRBs)*, observed as short flashes of high-energy photons followed by afterglows at lower photon energies (as discussed in §5.6). These afterglows can be used to study the first stars directly. To date, GRBs have been discovered by the *Swift* satellite out to $z = 9.4$, merely 540 million years after the Big Bang, and significantly earlier than the farthest known quasar ($z = 7.1$). It is already evident that GRB observations hold the promise of opening a new window into the infant Universe.

As discussed in §5.6, long-duration GRBs are believed to originate from the collapse of massive stars at the end of their lives (Figure 5.14). Since the very first stars were likely massive, they could have produced GRBs. If so, we may be able to see them one star at a time. The discovery of a GRB afterglow whose spectroscopy indicates a metal-poor gaseous environment, could potentially signal the first detection of a Population III star. The GRB redshift can be identified from the Lyman-α break in its otherwise power-law UV spectrum. A photometric detection can then be followed up with spectroscopy on a large telescope. Various space missions are currently proposed to discover GRB candidates at the highest possible redshifts.

In addition to individual source detections, GRBs are expected to reside in typical

small galaxies where massive stars form at those high redshifts. Once the transient GRB afterglow fades away, observers may search for the steady but weaker emission from its host galaxy. High-redshift GRBs may therefore serve as signposts of high-redshift galaxies which are otherwise too faint to be identified on their own. Importantly, GRBs are expected to trace the star formation history in a different way than galaxy surveys, since they can reside in galaxies below the survey detection threshold (although other biases, such as metallicity, may be important).

Moreover, standard light bulbs appear fainter with increasing redshift, but this is not the case with GRBs, because they are transient events that fade with time. When observing a burst at a constant *observed* time delay, we are able to see the source at an earlier time in its own frame. This is a simple consequence of time stretching due to the cosmological redshift. Since the bursts are brighter at earlier times, it turns out that detecting them at high redshifts is almost as feasible as finding them at low redshifts, when they are closer to us. It is a fortunate coincidence that the brightening associated with seeing the GRB at an intrinsically earlier time roughly compensates for the dimming associated with the increase in distance to the higher redshift, as illustrated by Figure 9.7.

9.3 LUMINOSITY AND MASS FUNCTIONS

The luminosity function (LF) of galaxies, $\phi(L)dL$, describes the number of galaxies per comoving volume within the luminosity bin between L and $L + dL$. It is the most fundamental observable quantity for galaxy surveys, and a great deal of effort has gone into measuring it in both the nearby and distant universe. Figure 9.8 shows measurements at $z = 4$–8 of the rest-frame ultraviolet galaxy luminosity function, with the most distant taken from the HUDF data. In this figure, the observed flux per unit frequency $(df/d\nu_{\rm obs})$ at an observed wavelength $\lambda_{\rm obs} = (c/\nu_{\rm obs})$ is translated to an equivalent AB magnitude using the relation,

$$M_{\rm AB} \equiv -2.5 \log_{10} \left[\frac{(df/d\nu_{\rm obs})}{\rm erg\ s^{-1}\ cm^{-2}\ Hz^{-1}} \right] - 48.6. \tag{9.1}$$

A popular fitting form for a wide range of galaxy surveys is provided by the *Schechter function*,

$$\phi(L) = \phi_\star \left(\frac{L}{L_\star} \right)^{-\alpha} \exp\left(-\frac{L}{L_\star} \right), \tag{9.2}$$

where the normalization $\phi_\star$ corresponds to the volume density at the characteristic luminosity $L_\star$, and α is the faint-end slope which controls the relative abundance of faint and bright ($L > L_\star$) galaxies. The total number density of galaxies is given by, $n_{\rm gal} = \int_0^\infty \phi(L)dL = \phi_\star \Gamma(\alpha + 1)$, and the total luminosity density is, $u_{\rm gal} = \int_0^\infty \phi(L) L dL = \phi_\star L_\star \Gamma(\alpha + 2)$, where Γ is the Gamma function. Note that at the faint end, $n_{\rm gal}$ diverges if $\alpha < -1$ and $u_{\rm gal}$ diverges if $\alpha < -2$. (In reality, the integrals converge anyway because there is a minimum luminosity for galaxies, set by a combination of the minimum halo mass for gas accretion and the minimum halo mass in which gas can cool.)

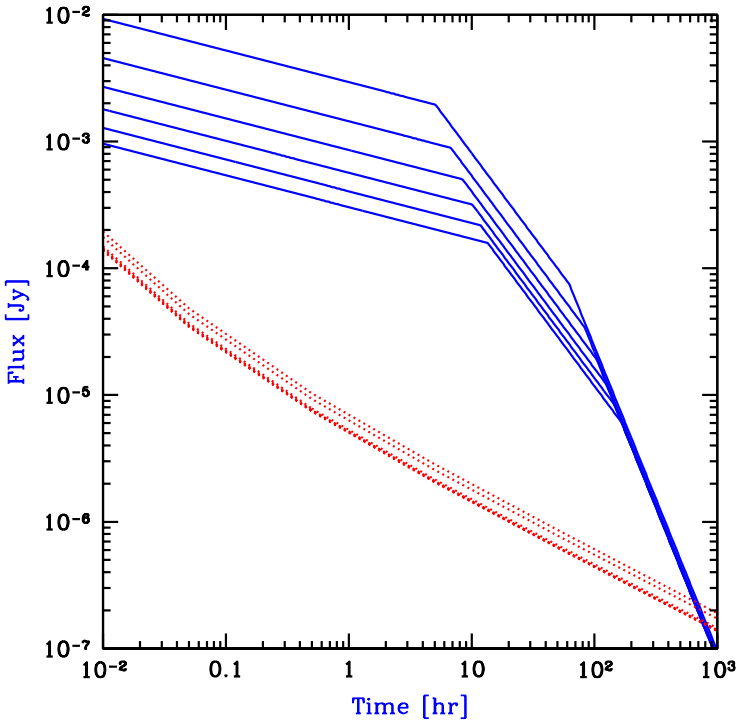

Figure 9.7 Detectability of high-redshift GRB afterglows as a function of time since the GRB explosion as measured by the observer. The GRB afterglow flux (in Jy) is shown at the redshifted Lyman-α wavelength (solid curves). Also shown (dotted curves) is a crude estimate for the spectroscopic detection threshold of *JWST*, assuming an exposure time equal to 20% of the time since the GRB explosion. Each set of curves spans a sequence of redshifts: $z = 5, 7, 9, 11, 13, 15$, respectively (from top to bottom). Figure credit: Barkana, R., & Loeb, A. *Astrophys. J.* **601**, 64 (2004).

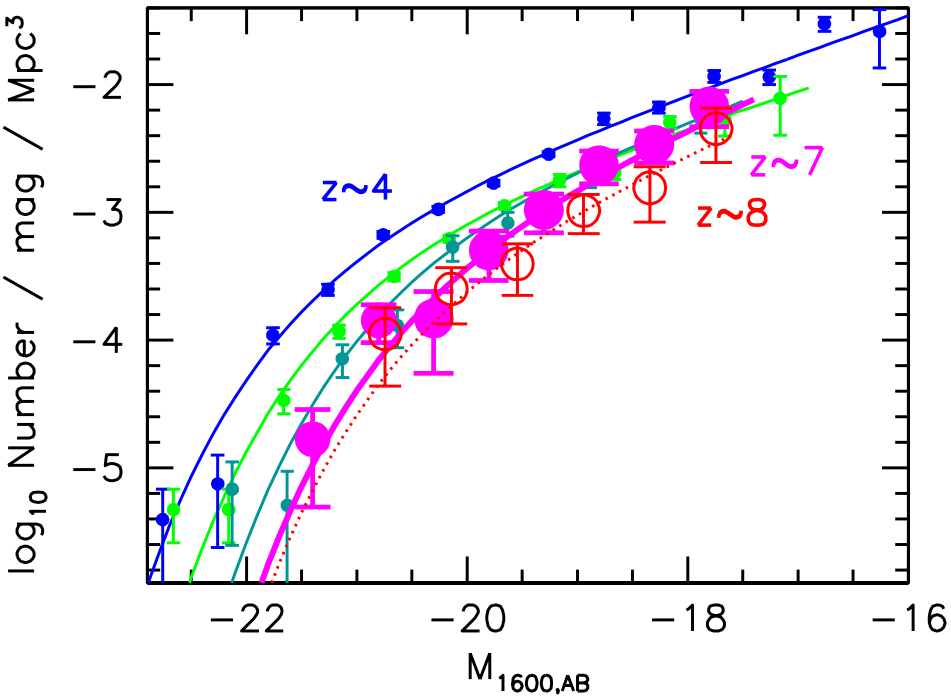

Figure 9.8 Rest-frame ultraviolet luminosity functions derived for galaxies at $z \sim 7$ (large filled circles) and $z \sim 8$ (large open circles) compared to lower redshift data (from $z = 6$ to $z = 4$). The vertical axis gives the number of galaxies per comoving Mpc^3 per AB magnitude at a rest-frame wavelength of 1600Å, as a function of this magnitude on the horizontal axis. Note the sharp decline in the number density of bright galaxies with redshift and tentative evidence for a steepening faint-end slope. Figure credit: Bouwens, R.. et al. *Astrophys. J.*, in press (2011).

The curves in Figure 9.8 show fits of this form to the data; clearly this simple, empirical structure does an excellent job of matching the observations. Three points emerge from these fits: *(1)* the characteristic luminosity $L_\star$ declines toward higher redshift; *(2)* the space density of galaxies at $L_\star$ also decreases; and *(3)* the faint-end slope may steepen at $z > 7$ (though the evidence for this is still tentative). In light of typical models for structure formation, in which these galaxies are associated with dark matter halos, this is hardly surprising: at higher redshifts, fewer halos have formed, and in any hierarchical model those that have formed are preferentially smaller. The interesting physics involves the mapping from the halo mass function to luminous baryons, which we discuss below.

The particular physical insight provided by a galaxy survey depends upon the selection technique and waveband used. In general, rest-frame ultraviolet measurements (such as those shown in Fig. 9.8) depend exclusively on hot stars able to produce the observed UV photons. Because these high-mass stars are short-lived, the UV luminosity is tied to the star formation rate (SFR) of the galaxies, although there is an uncertain correction that depends on the IMF of stars. In fact, there are several ways to estimate SFRs from other measurements as well (the following conversions assume a standard Salpeter IMF):

- **The rest-frame UV continuum** (1250–1500Å) - provides a direct measure of the abundance of high-mass $> 5M_\odot$ main-sequence stars. Since these stars are short lived, with a typical lifetime $\sim 2 \times 10^8$ yr$(m_\star/5M_\odot)^{-2.5}$, they provide a good measure of the star formation rate, with

$$SFR \approx 1.4 \left(\frac{L_\nu}{10^{28} \text{ erg s}^{-1} \text{ Hz}^{-1}} \right) M_\odot \text{ yr}^{-1}. \tag{9.3}$$

  The primary uncertainty in this determination is extinction via dust, though that can be estimated from the spectra or from other observations.

- **Nebular emission lines**, such as Hα and [OII], measure the combined luminosity of gas clouds which are photo-ionized by very massive stars ($> 10M_\odot$). Dust extinction can be evaluated from higher-order Balmer lines, but this estimator is highly sensitive to the assumed IMF. For the Milky-Way IMF,

$$SFR \approx 0.8 \left(\frac{L(\text{H}\alpha)}{10^{41} \text{ erg s}^{-1}} \right) M_\odot \text{ yr}^{-1}, \tag{9.4}$$

  and

$$SFR \approx 1.4 \left(\frac{L([\text{OII}])}{10^{41} \text{ erg s}^{-1}} \right) M_\odot \text{ yr}^{-1}. \tag{9.5}$$

- **Far-infrared emission** (10–300μm) - measures the total emission from dust heated by young stars,

$$SFR \approx 0.45 \left(\frac{L(\text{FIR})}{10^{43} \text{ erg s}^{-1}} \right) M_\odot \text{ yr}^{-1}. \tag{9.6}$$

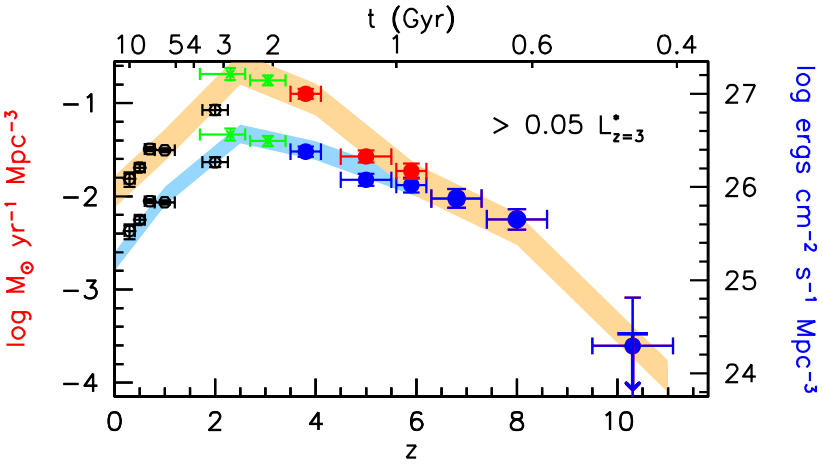

Figure 9.9 The star formation rate density (left vertical axis) or luminosity density (right vertical axis) as functions of redshift (lower horizontal axis) and cosmic time (upper horizontal axis), for galaxies brighter than an AB magnitude of -17.7 (corresponding to $0.05L_\star$ at $z = 3$). The conversion from observed UV luminosity to star formation rate assumed a Salpeter IMF for the stars. The upper curves includes dust correction based on estimated spectral slopes of the observed UV continuum. Figure credit: Bouwens, R., et al., preprint http://arxiv.org/pdf/1006.4360v4 (2010).

- **Radio emission**, for example at a frequency of 1.4 GHz – measures the synchrotron emission from relativistic electrons produced in supernova remnants. The supernova rate is related to the "instantaneous" production rate of massive stars ($> 8M_\odot$), because these have a short lifetime, giving on timescales longer than $\sim 10^8$ yr,

$$SFR \approx 1.1 \left(\frac{L_\nu(1.4\mathrm{GHz})}{10^{28}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}} \right)\ M_\odot\ \mathrm{yr}^{-1}. \qquad (9.7)$$

Integration of the luminosity function of galaxies with a kernel that measures their star formation rate yields the star formation rate per comoving volume in the Universe. Figure 9.9 shows a recent determination of this rate based on the UV luminosity function as a function of redshift for all galaxies brighter than $0.05L_\star$ at $z = 3$ (corresponding to an AB magnitude of -17.7). Most of these measurements are made from UV data, so the correction for dust extinction is particularly important (shown by the upper set of measurements here). Also, one must note that this is a *lower* limit to the true star formation rate density, because it ignores feeble galaxies below the detection threshold.

One obvious omission from our list of star formation rate indicators is the Lyman-α line, which as we discussed is useful in detecting high-redshift galaxies. However, as we will see in chapter 10, the interpretation of this emission line is fraught

with uncertainties about the galaxy's dust content, ISM structure, outflow properties, and environment. Therefore, the Lyman-α line is not a good star formation rate indicator. However, one can still construct a luminosity function of emission in this line; Figure 9.10 shows recent determinations in the redshift range of $z = 3$–6.6. In contrast to the luminosity function of photometrically-selected LBGs, LAEs do not appear to change in comoving number density between $z = 3$–5.7, although their density appears to decline rapidly beyond that. This may be an indication of changes in the galaxy environments – and possibly reionization – though that interpretation is very controversial.

Meanwhile, the mass budget of stars at $z \sim 5$–6 can be inferred from their rest frame optical and near-infrared luminosities, which are much closer to measuring the total stellar content than ultraviolet light, because low-mass stars emit in these bands. Measuring this total stellar content is more physically interesting than the star formation rate density because the cumulative density provides a census of stars that were made in faint galaxies below the detection threshold and only later incorporated into detectable galaxies. Moreover, it provides some information on the *past* history of star formation (though still subject to uncertainty with the IMF). Figure 9.11 shows some recent measurements of the growth of the total stellar mass density in the Universe. Note in particular that only a small fraction of the stars present at $z < 2$ formed at $z > 6$, though this is subject to an unknown correction from undetected galaxies.

9.4 THE STATISTICS OF GALAXY SURVEYS

Measurements of the statistical properties of galaxies are challenging, and in this section we will discuss strategies to constrain their properties, including "one-point functions" like the luminosity or stellar mass functions as well as spatial correlations. The former generally provide insight into the baryonic physics of galaxy formation – how dark matter halos accrete gas and transform it into stars – while clustering provide insight into the dark matter halos themselves.

9.4.1 Estimates of Galaxy Clustering

We have already described our primary tool for understanding the spatial distribution of galaxies, the power spectrum, in §3.7.5, where we developed it through the halo model. In this representation, it contains two terms: the *two-halo* term, which describes the correlations between distinct dark matter halos, and the *one-halo* term, which describes the distribution of galaxies within an individual dark matter halo. At high redshifts, where halos are small and probably host only one "galaxy" (at least as we would define them observationally), the latter likely disappears in most halos. Thus, in the language of §3.7.5, we will adopt $\langle N|m\rangle = 1$ for all halos above a minimum mass set by accretion or feedback, $m_{\rm min}$, and zero otherwise.

The key additional input that we need is some way of mapping the observed luminosity (in some photometric band or emission line) or other observable to halo

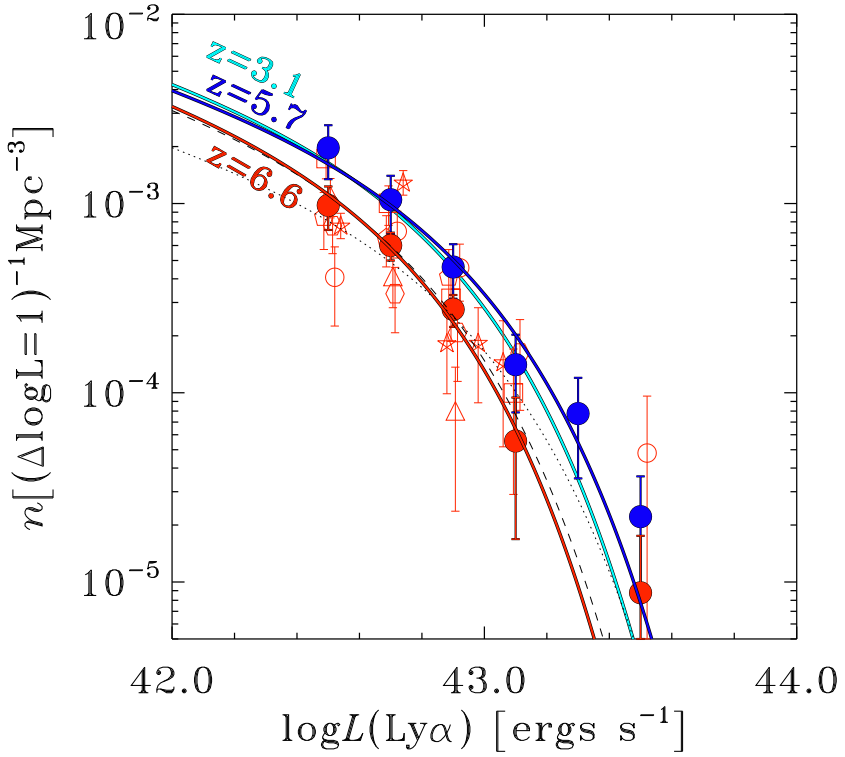

Figure 9.10 Luminosity function of LAEs from $z = 3$–6.6. The light solid circles show the measured luminosity function at $z = 6.6$, while the lighter solid circles show the same for $z = 5.7$. The solid line show Schechter function fits to these as well as the best fit at $z = 3.1$ (lightest curve). The LAE density drops substantially from $z = 5.7$ to $z = 6.6$, much faster than that of LBGs. Finally, the open symbols show the number densities measured in the five sub-fields of the $z = 6.6$ survey, illustrating the substantial variance between fields. Figure credit: Ouchi, M. et al. *Astrophys. J.* **723**, 869 (2010).

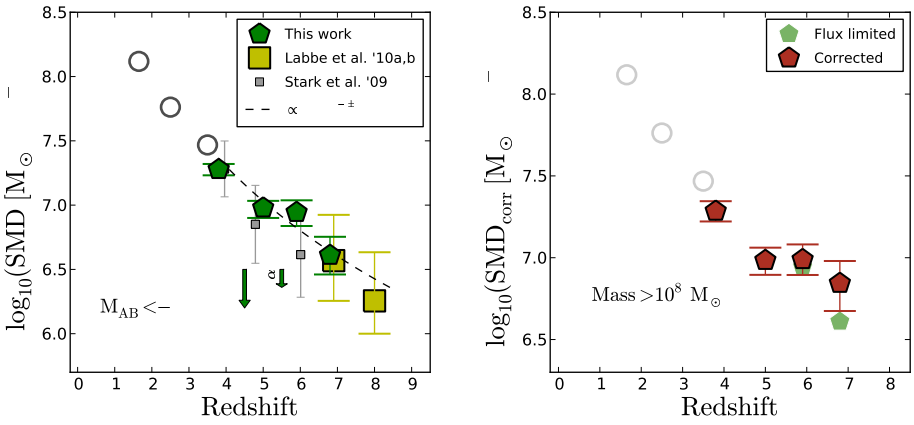

Figure 9.11 Stellar mass density (SMD) evolution over redshift in galaxies brighter than $M_{\rm AB} = -18$ at a rest-frame wavelength of 1500Å (*left*) and with stellar mass $M_\star > 10^8\ M_\odot$ (*right*) from several different measurements. All panels assume a Salpeter IMF and $Z = 0.2\ Z_\odot$; the line labeled "Chabrier" shows the effect of assuming a different IMF. The other line (labeled "Hα") shows the effect of systematic contamination from line emission. A minimum stellar density $\sim 1.7 \times 10^6 f_{\rm esc}^{-1}\ M_\odot$ Mpc^{-3} of Population III stars (corresponding to $\Omega_\star \sim 1.25 \times 10^{-5} f_{\rm esc}^{-1}$) is required to produce one ionizing photon per hydrogen atom. Figure credit: Gonzalez, V. et al. *ApJL* **735**, L34 (2011).

mass. One framework for doing so is the *conditional luminosity function*, involving a function $\phi(L|m)$ which parameterizes the luminosity distribution of halos with a given mass. We will describe the complex physics behind this function in the following sections; for now we assume that it can be constructed from some theoretical or empirical arguments. As a crude example, however, we could suppose that galaxies are luminous for a fraction $f_{\rm duty}$ of the time but that there exists a one-to-one relationship between luminosity and mass, $L(m)$, while they are "on." In that case,

$$\phi(L|m) = (1 - f_{\rm duty})\delta(0) + f_{\rm duty}\delta(L[m]). \tag{9.8}$$

Assuming only one galaxy per halo, and given a minimum observable luminosity $L_{\rm min}$, the predicted galaxy power spectrum will be

$$P^{\rm gal}(k,z) = P_{\rm lin}(k,z)\left[\int dm \frac{f(>L_{\rm min}|m)}{\bar{n}_{\rm obs}} n(m,z) b_{\rm eff}(k|m,z)\right]^2, \tag{9.9}$$

where

$$f(>L_{\rm min}|m) = \int_{L_{\rm min}}^{\infty} dL\, \phi(L|m). \tag{9.10}$$

Comparison to equation (3.85) shows that we have dropped the halo profile (under the assumption that each halo contains only one galaxy) and replaced $\langle N|m\rangle$ with $f(>L_{\rm min}|m)$, which is the probability that a halo of mass m hosts an observable galaxy – clearly, these are two different ways of expressing the occupation fraction of dark matter halos. One can then define a mean bias for this galaxy sample, $\bar{b}_{\rm eff}(k)$ by averaging over all the observable galaxies, so that $P_{\rm gal,obs} \approx \bar{b}_{\rm eff}(k)^2 P_{\rm lin}(k,z)$; in the limit of linear fluctuations, this mean bias is independent of scale and can be predicted using the excursion set formalism in equation (3.55).

Because this effective bias depends on the underlying mass of the galaxy halos – a property of the population that is otherwise nearly impossible to measure – the galaxy power spectrum is of fundamental importance. We will also see in the next section that it can shed important light on reionization as well. We will therefore next describe how it can be measured and the errors that appear when doing so.

Let us suppose that we have a survey over some finite volume V. For now we will assume that the three-dimensional locations of the galaxies are known through some spectroscopic survey. Let us define the galaxy overdensity for a mode k_i through the Fourier transform of the galaxy density field,

$$\delta(\mathbf{k}_i) = \int \frac{d^3x}{V} W(\mathbf{x})\delta_g(\mathbf{x}) e^{i\mathbf{k}_i\cdot\mathbf{x}}, \tag{9.11}$$

where $W(\mathbf{x})$ is the *window function*, which is non-zero only inside the survey volume and is normalized so that $\int d^3x W(\mathbf{x}) = V$, the total survey volume.[iii] We then write

$$\langle \delta(\mathbf{k}_i)\delta(\mathbf{k}_j)\rangle = \mathbf{C}^S_{ij} + \mathbf{C}^N_{ij}, \tag{9.12}$$

[iii] In practice, we can incorporate any other selection function, such as the likelihood of detecting a galaxy at a particular redshift given some photometric selection criterion, into the window function. Such practical considerations make the analysis more complex but do not change the basic methodology we describe.

where $\mathbf{C}^S$ is the *signal covariance matrix* and $\mathbf{C}^N$ is the *noise covariance matrix*. Here the angular brackets denote an average over the density modes in the Universe.

Substituting equation (9.11) into $\langle\delta(\mathbf{k}_i)\delta(\mathbf{k}_j)\rangle$, this average operates on the galaxy overdensity terms; from equation (**??**), this gives the correlation function of the galaxy field:

$$\mathbf{C}^S_{ij} = \frac{1}{V^2}\int d^3x\, d^3x'\, W(\mathbf{x})W(\mathbf{x}')\xi^{\rm gal}(\mathbf{x}-\mathbf{x}')e^{i\mathbf{k}_i\cdot\mathbf{x}}e^{-i\mathbf{k}_j\cdot\mathbf{x}'}. \tag{9.13}$$

However, the correlation function is simply the Fourier transform of the power spectrum (see eq. 2.15), so we can write

$$\mathbf{C}^S_{ij} = \int \frac{d^3k}{(2\pi)^3}P^{\rm gal}(k)\frac{\tilde{W}(\mathbf{k}_i-\mathbf{k})\tilde{W}^*(\mathbf{k}_j-\mathbf{k})}{V^2}, \tag{9.14}$$

where $\tilde{W}(\mathbf{k}')$ is the Fourier transform of the window function. Comparison to equation (**??**) shows that this closely related to the variance in the fluctuations over the volume of the survey.

To gain some intuition for this expression, let us consider some concrete choices for the window function. First, suppose that we observe a spherical volume of radius R around some central point $\mathbf{x}_0$.[iv] Then

$$\tilde{W}(\mathbf{k}_i-\mathbf{k}) = \int_{|\mathbf{x}-\mathbf{x}_0|<R} d^3x\, e^{i(\mathbf{k}_i-\mathbf{k})\cdot\mathbf{x}}. \tag{9.15}$$

In the limit that $R\to\infty$, this is proportional to a Dirac delta function, so we would have

$$\mathbf{C}^S_{ij} \approx (2\pi)^3 P(k_i)\delta^D(\mathbf{k}_i-\mathbf{k}_j), \tag{9.16}$$

which matches our original definition of the power spectrum (eq. **??**). A finite R broadens the delta function, so that the Fourier transform has a non-zero width $\sim(2\pi)/R$. This means that the measured signal will be a weighted average of all modes with $|\mathbf{k}-\mathbf{k}_i| < (2\pi)/R$. Modes with wavelengths larger than the survey volume will be unobservable – they have such small k as to be washed out; those with $k \gg 1/R$ will be essentially unaffected.

At least for the time being, a more realistic survey geometry is a "pencil-beam:" a narrow angular region (a few arcminutes across for HST or JWST) with a large depth in the radial direction, corresponding perhaps to Lyman-break selection within a particular filter set. In that case, the volume may reasonably be approximated as a long with radial depth Δz and transverse widths Δx and Δy, such that $\Delta z \gg \Delta x,\ \Delta y$. For a rectangular box, the window function is

$$\tilde{W}(\mathbf{k}) = W_{\Delta x}(k_x)W_{\Delta y}(k_y)W_{\Delta z}(k_z), \tag{9.17}$$

with (k_x, k_y, k_z) the Cartesian components of the wavevector, and

$$W_{\Delta x}(k_x) = \frac{\sin(k_x\Delta x/2)}{(k_x\Delta x)/2}, \tag{9.18}$$

[iv] Choosing the central point to be the observer would correspond to a volume-limited sample of galaxies around us. However, for high-redshifts, the center of the survey would naturally lie at some distant point.

and similarly for $W_{\Delta y}$ and $W_{\Delta z}$. This function also is ~ 1 for $k_x \ll \pi/\Delta x$ and falls off at $k_x \sim \pi/\Delta x$. The anisotropy of the window means that the mode sampling depends on their orientations. Modes transverse to the line of sight must have $k_{x,y} \ll 2\pi/\Delta x$ in order to be sampled cleanly, but modes along the line of sight must only have $k_z \ll 1/\Delta z$. Even these modes, however, are subject to aliasing from short-wavelength transverse modes, similarly to the Lyman-α forest power spectrum discussed in §4.3.4.

The noise term $\mathbf{C}^N$ arises from the finite number of galaxies. This so-called *shot noise* term is inevitable in any experiment that samples a discrete population of objects. Let us assume that the number of galaxies within a given volume obeys Poisson statistics with the mean expected count $\bar{N}$ determined by the underlying density field. The probability of finding N galaxies in a region is then $p(N) = \bar{N}^N e^{-N}/N!$, with $\langle N \rangle = \bar{N}$ and $\langle N^2 \rangle = \bar{N}(\bar{N}+1)$. For this discrete shot noise component, the average in equation (9.12) becomes $\langle \delta_i \delta_j \rangle = \bar{N}^{-2} \langle (N_i - \bar{N})(N_j - \bar{N}) \rangle = \bar{N}^{-1}$ if $i = j$ and zero otherwise. This expression replaces the power spectrum in equation (**??**). Finally, we assume that we can choose regions sufficiently small so that each one is either empty or contains at most one galaxy; in that case $\bar{N} = \bar{n}$, the galaxy number density. Finally, by analogy to equation (**??**), we define the *shot noise power spectrum* as $P^{\rm shot}(k) = 1/\bar{n}$, or

$$\Delta^2_{\rm shot}(k) = \frac{k^3}{2\pi^2 \bar{n}}. \tag{9.19}$$

This is an inevitable source of noise in any galaxy survey; fortunately, provided one has a good estimate for $\bar{n}$ it can be accurately removed. Shot noise therefore only poses a significant problem when $\bar{n}$ is small, for example if the survey targets extremely bright galaxies with $L \gg L_\star$ which are rare.

The power spectrum is by far the most common measure of clustering, owing to the relative ease with which it can be observed. However, it only measures the variance (as a function of scale) of the underlying distribution; higher-order correlations, like skewness, must be measured in other ways. A particularly simple approach to test for these is with *counts-in-cells*, in which one divides the survey volume into small cells and examines the distribution of galaxy counts in the cells.

Another complication arises if the radial locations of the galaxies are not available, for example if the galaxies are found through the Lyman-break technique without precise redshifts. In that case, clustering can still be measured along the plane of the sky. This *angular correlation function* (or its counterpart the angular power spectrum) was traditionally the best measure of clustering, even at low redshifts. Intuitively, the angular correlation function will simply be the projection of the three-dimensional form onto the plane of the sky. For small angular separations, this is easy to do; we will discuss it in more detail in §12.2.

We also note that, whenever redshift is used as a proxy for distance (as in a spectroscopic galaxy survey), peculiar velocities in the galaxy field will distort the redshift-distance mapping. Fortunately, the velocity effects can be isolated, because they do not affect positions across the plane of the sky: we therefore expect a difference in the clustering measured along the line of sight and along the plane of the sky. Because these peculiar velocities themselves trace the underlying matter distri-

bution, the corresponding *redshift-space distortions* can themselves be particularly useful. We will discuss them farther in §11.5.1.

9.4.2 Measuring the Luminosity Function

In addition to its intrinsic interest as a measure of halo mass, clustering also affects the statistical uncertainty in number counts of galaxies within surveys of limited volume, the so-called *cosmic variance*. This is crucial to understand for estimates of luminosity and stellar mass functions, because it determines the precision of such measurements. By analogy to equation (**??**), the fractional variance in an estimate of the galaxy number counts is the integral of the signal and noise power spectra over all k-modes, weighted by the survey window function:

$$\frac{\sigma^2_{\rm tot}}{\langle N\rangle^2} = \int \frac{dk}{k}\left[\Delta^2_{\rm gal}(k) + \Delta^2_{\rm noise}(k)\right]\frac{|W(\mathbf{k})|^2}{V^2}. \tag{9.20}$$

The cosmic (or sample) variance, which is the first term on the right hand side of equation (**??**), results from the survey field sometimes lying in a region of high galaxy density and sometimes lying in an under-dense region or a void.

Figure 9.12 compares the contributions from cosmic variance and shot noise as calculated by linear theory for a mock survey as a function of its opening angle, $\theta = a_x/\chi(z)$. This plot can be used to estimate the effectiveness of future surveys with large fields of view. Here we have used a simple model to assign luminosities to dark matter halos, taking $f_{\rm duty} = 0.25$ and a star formation efficiency $f_\star = 0.16$. Note how the shot noise is only important on small scales, even though the fluctuations from gravitational clustering also decrease with the opening angle of the survey.

According to linear theory, the probability distribution of the count of galaxies is a Gaussian with variance given by the sum of the cosmic and Poisson components, so the power spectrum provides a complete representation. However, nonlinear evolution in the matter field induces non-Gaussian structure; because bright high-redshift galaxies are so rare, these nonlinearities can have important effects. Figure 9.13 shows this in the context of a pencil-beam survey of galaxies (with a $3.4' \times 3.4'$ field of view, as for the HUDF) in the redshift range $z = 6 - 8$. When compared to numerical simulations, the galaxy count statistics are well approximated by the linear-theory expressions at the low luminosity limits.

However, for brighter galaxies linear theory begins to fail. The upper solid curve in the top panel of Figure **??** shows the variance calculated from numerical simulations that include nonlinear evolution. These are larger than the analytic prediction (shown by the lower solid curve) for halo masses $M > 10^{10}\ M_\odot$. With the simulations, one can investigate how this happens. The *skewness* is

$$s_3 = \frac{\left\langle (N - \langle N\rangle)^3\right\rangle}{\left(\sigma^2\right)^{3/2}}. \tag{9.21}$$

The skewness as a function of minimum luminosity is presented in the bottom panel of Figure 9.13. It is large at $M_h > 10^{10}\ M_\odot$ (the seemingly large amplitude variations in the skewness at low luminosity for $z = 6$–8 are due to small numerical

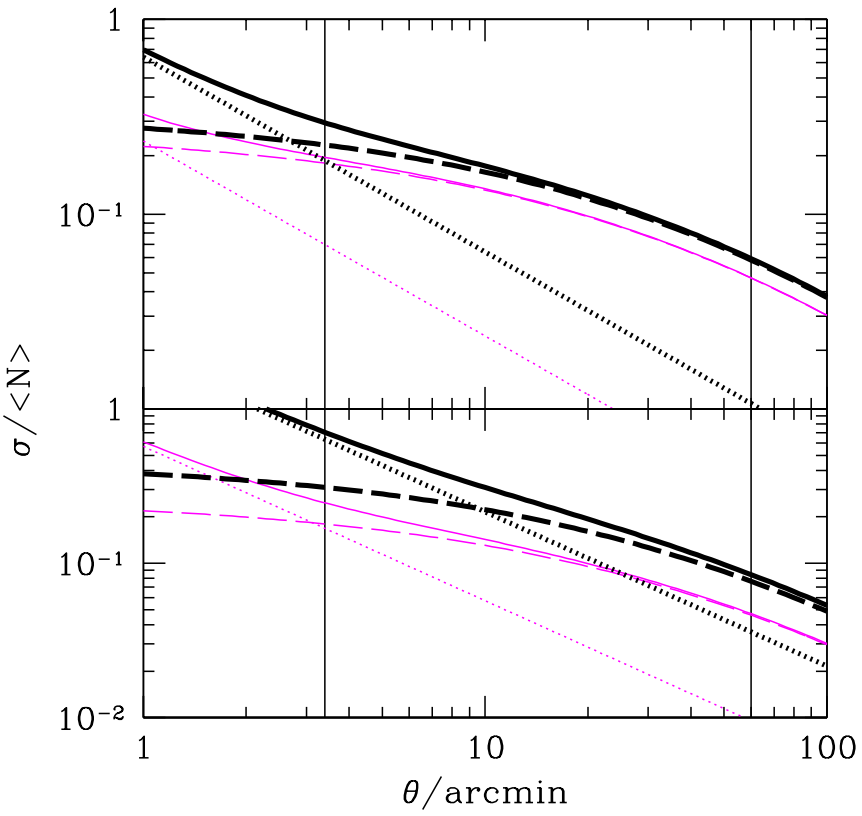

Figure 9.12 The theoretically predicted contributions to the total variance (equation **??**; solid lines) in Lyman-break galaxy dropout surveys as a sum of cosmic variance (dashed lines) and Poisson shot noise (dotted lines) contributions. The top and bottom panels show results for surveys extending from $z = 6$–8 and $z = 8$–10, respectively. Thin lines assume a luminosity threshold of $z_{850,\mathrm{AB}}$=29, while for thick ones, the cut is at $z_{850,\mathrm{AB}}$=27. Figure credit: Munoz, J., Trac, H., & Loeb, A. *Mon. Not. R. Astron. Soc.*, **405**, 2001 (2010).

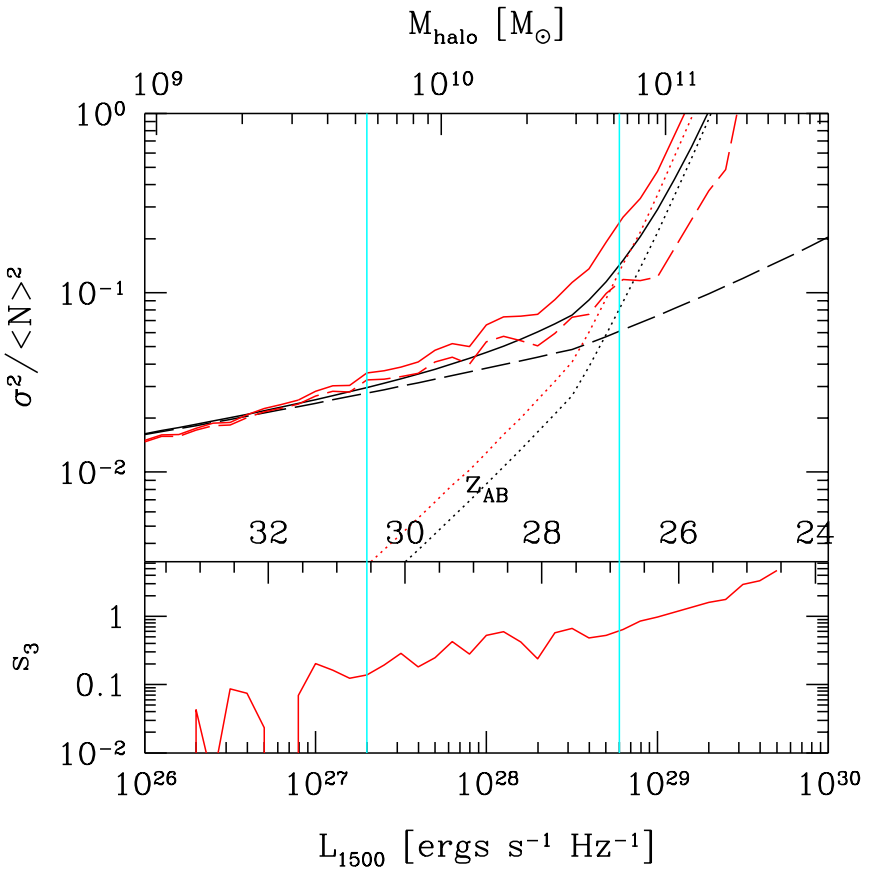

Figure 9.13 *Upper panel:* Predicted relative contributions to the fractional variance in the number counts of galaxies as a function of UV luminosity at an emission wavelength of 1500Åwithin a Lyman-break dropout survey spanning the redshift interval $z = 6$–8 with a $3.4' \times 3.4'$ field-of-view (matching HUDF and approximately that of JWST). Solid lines show the total variance, while long-dashed and dotted lines show the contributions from cosmic variance and Poisson noise, respectively. The upper curves show the results from numerical simulations, while the lower curves were calculated analytically based on linear perturbation theory. Vertical lines bracket the region where the variance is higher than expected due to the skewness of the full count probability distribution but is not Poisson-dominated. The middle and top horizontal axes translate the monochromatic luminosity to z-band AB magnitude and host halo mass, respectively. *Lower panel:* Skewness of the full galaxy count probability distribution calculated from a numerical simulation based on equation (9.21). Figure credit: Munoz, J., Trac, H., & Loeb, A. *Mon. Not. R. Astron. Soc.*, **405**, 2001 (2010).

fluctuations around the near-zero skewness from numerical simulations, plotted on a log scale). The numerical simulations indicate that the probability distribution of massive halos (and hence presumably bright galaxies) has a non-Gaussian shape. Deviations between the analytic and simulation values of the sample variance grow when the skewness becomes significant. This behavior is a manifestation of non-linear clustering on the small scales probed by the narrowness of the survey skewer.

9.4.3 Measuring the Galaxy Power Spectrum

We have now shown how to estimate the galaxy power spectrum and how its fluctuations affect number counts of galaxies (and hence the luminosity function). As a final step, let us briefly discuss the errors on a measurement of the power spectrum itself: how large of a survey is necessary in order to reliably measure the clustering of a galaxy sample?

Given that real galaxy surveys have complex selection functions, the best way to answer this question is of course with detailed simulations of the survey strategy. The next best way is with the **Fisher information matrix**, which provides a robust lower limit to the errors on a given set of parameters in any experiment. We will consider this latter approach here. Suppose one wishes to measure the amplitude of the power spectrum over a range of wavenumbers $(k, k + \Delta k)$ in a survey of volume V (these are known as *band powers*). Ignoring boundary effects from the finite survey volume, the minimal error on the band powers is

$$\frac{\Delta P(k)}{P(k)} = (2\pi)\sqrt{\frac{2}{k^2 \Delta k \Delta\mu V}}\left[\frac{1+\bar{n}P(k)}{\bar{n}P(k)}\right]. \tag{9.22}$$

This expression is straightforward to understand. Recall that the power spectrum quantifies the variance in the density field amongst a set of modes. Suppose we have N independent estimates of these mode amplitudes. From elementary statistics, the mean squared error on an estimate of the variance from this dataset will be $(\Delta P)^2 \sim \sigma^2/N$, where σ is the variance of the measurements: in our case, $P + 1/\bar{n}$. We thus only need determine the number of independent samples of the density field in a given power spectrum bin. First, let us write the Fourier space volume of in a binned measurement of the power spectrum as $2\pi k^2 \Delta k \Delta\mu$, where μ is the cosine of the angle between the bin's central wavevector and the line of sight. (As mentioned briefly in §9.4.1 and more extensively in §11.5.1, peculiar velocities induce an anisotropy with respect to μ. For a crude measurement, however, we can average over all modes with a single amplitude, so that $\Delta\mu = 2$ and the volume corresponds to a spherical shell in k-space.)

The final question is how many samples lie within this Fourier space volume. Recall from §9.4.1 that the finite survey volume mixes all modes closer together than $\sim (2\pi)^3/V$. However, the reality of the density field imposes a constraint on its Fourier transform, relating pairs of modes with $\mathbf{k}$ and $-\mathbf{k}$ to each other. Thus, the number of *independent* samples is $N \approx 2\pi k^2 \Delta k \Delta\mu \times [V/(2\pi)^3] \times 1/2$. The prefactor in equation (9.22) is simply $1/\sqrt{N}$.

This approach provides a reasonable estimate for the volume required to measure galaxy bias from a survey. In the regime where shot noise is unimportant

(i.e., $P \gg \bar{n}^{-1}$), a measurement with 10% precision requires a volume of $\sim 10^4 (k/0.1\,\mathrm{Mpc}^{-1})^{-3}\,\mathrm{Mpc}^3$. High-$k$ modes evidently do not require particularly large volumes, but related surveys run into shot noise limitations unless they go very deep; even the faintest HUDF galaxies have $\bar{n} \sim 0.01\,\mathrm{Mpc}^{-3}$ (see Fig. 9.8). Shot noise compromises modes with $\bar{n}P < 1$. In that case, it is generally advantageous to construct a deeper, rather than wider, survey.

9.5 THE PHYSICS OF GALAXY EVOLUTION

Attempts to reproduce the evolution in the luminosity function of galaxies over cosmic time require various mechanisms of feedback from reionization, supernovae, and gas accretion onto a central black hole. A comprehensive understanding of the physical details of these feedback processes (often used in semi-analytic models of the luminosity function as a function of redshift) is lacking. Nevertheless, we can at least identify the important processes that drive the evolution of galaxies. In this section, we will briefly describe these ingredients, dedicating special attention to how they might affect models and observations at high redshifts.

The starting point for understanding the abundance, clustering, and other properties of galaxies is the dark matter halo distribution, $n(m)$. We wish to understand the mapping from halo mass to luminosity (in many different bands or lines), stellar mass, metallicity, star formation history, velocity distribution, and any other physical properties of interest. Of course, there is nothing to demand that this mapping is one-to-one, or even that these physical properties depend exclusively on halo mass, as they could also depend on the halo's larger-scale environment, mass accretion history, etc. The challenge of research on galaxy formation and evolution is to understand which factors are most important, and how all of them interact to produce the objects we observe.

On the coarsest scale, galaxies are machines that transform accreted material (whether acquired through smooth accretion or mergers) into stars and black holes. The crucial complication is feedback, which can both prevent gas from accreting onto a halo in the first place and expel material that is already present (preventing it from forming stars, or providing potential fuel for later accretion episodes). Because this feedback is generated on the smallest scales (through stars or black holes), understanding galaxy evolution requires a model spanning a large range of physical scales and processes.

Theoretical astrophysicists examine many of these problems individually (and hence generally in isolation from each other). Their results inform coarser models – including both "semi-analytic" approaches that rely on relatively simple models for the many processes involved, as well as numerical simulations (which rarely span the required dynamic range and so also contain simple prescriptions for at least some of the processes). For the sake of brevity, we will follow the latter approach and aim only to parameterize the important processes and suggest some intuition for the underlying physical processes. This is by no means a comprehensive treatment but should give a flavor for the "less" exotic processes that affect galaxy evolution even at the present day and that are described in many other textbooks (see *Further*

Reading for related resources).

The very simplest model involves two free parameters: *(i)* the fraction of baryons converted into stars within a host halo, $f_\star$; and *(ii)* duty cycle of vigorous star formation activity during which the host halo is luminous and the stars are formed. This model defines the star formation timescale, $t_\star$, as the product of the star formation duty cycle, $\epsilon_{\rm DC}$, and the cosmic time, $t_{\rm H}(z) = 2/3H(z)$ at the redshift of interest z. The star formation rate $\dot{M}_\star$ is then related to halo mass M as follows

$$\dot{M}_\star(M) = \frac{f_\star \times (\Omega_b/\Omega_m) \times M}{t_\star}. \tag{9.23}$$

Within the context of this simplest model, the physics of galaxies determines the values of $f_\star$ and $t_\star$ (which may depend on halo mass as well, and presumably have some scatter between different halos at the same mass). More complex models are obviously necessary to track more detailed properties of the sources as well. Some of these ingredients are described below.

9.5.1 Gas Accretion

The fuel for star and black hole formation inside galaxies is provided by gas accretion onto the halo, either in a relatively slow, steady mode or in a stochastic, "merger" mode. The first is relatively easy to model. First consider a spherical system. Provided that the gas accretes supersonically – as it will if the halo has a virial temperature $T_{\rm vir}$ larger than the IGM temperature – we would generically expect an accretion shock to form, at a radius comparable to the virial radius of the halo. However, such a shock is only stable if the hot gas behind it can support the shock. If, on the other hand, this gas can cool rapidly, the shock will sink inward until it stabilizes very near the galaxy when the sound-crossing time $t_{\rm sc} \sim r/c_s$ becomes smaller than the radiative cooling time per proton,

$$t_{\rm cool} = \frac{3}{2}\frac{1}{\mu m_p}\frac{kT}{\rho(r)\Lambda(T,Z)}, \tag{9.24}$$

where $\rho^2\Lambda$ is the radiative cooling rate per unit volume. It is dominated by Compton cooling (at very high redshifts) and atomic line transitions, and it therefore depends sensitively on the metallicity and temperature of the gas. In more detail, the condition for virialization shock stability is

$$\frac{\rho r \Lambda(T,Z)}{u^3\mu^2 m_H^2} < 0.0126, \tag{9.25}$$

where T is the postshock value and u is the infall velocity.

The transition between these rapid and slow cooling regimes depends on the halo mass and redshift. Crudely, halos with masses above this critical threshold $M_{\rm cool}$ will have hot "atmospheres" that cool slowly. Gas will accrete onto these atmospheres rather than the galaxy; in that case, the rate of accretion onto the galaxy itself will be limited not by cosmological processes but by cooling within the atmosphere. On the other hand, halos with mass $M < M_{\rm cool}$ will be limited only by the cosmological infall rate (and feedback from the galaxy itself; see below). This critical threshold occurs at $\sim 10^{11}\ M_\odot$ for gas with primordial composition,

or $\sim 10^{12}\ M_\odot$ for gas with solar composition, with only a mild dependence on redshift (the left-hand side of equation (9.25) is $\propto \rho r_{\rm vir}\Lambda/u^3 \propto (1+z)^{1/2}\Lambda$). This is sufficiently large that most high-redshift galaxies will be fed through the rapid-cooling regime and so be limited only by the rate of cosmological infall.

Recently, high-resolution simulations have shown that the filamentary geometry of the cosmic web changes this picture slightly; the "rapid-cooling" regime is generally fed by accretion along filaments that reaches the galaxy's star-forming region without any shocks at all. These *cold flows* provide the primary fuel supply for small galaxies, but the filamentary structures can persist in larger galaxies as well. The transition between these two cooling regimes is therefore not an abrupt one.

The overall growth of halo mass can be tracked with analytic arguments and simulations. In particular, the extended Press-Schechter formalism described in §3.5.2 provides a mechanism to estimate this, which matches numerical simulations reasonably well. In the standard cosmology, this approach yields

$$\frac{\dot{M}/M}{H(z)} \approx 2.3 \left(\frac{M}{10^{10}\ M_\odot}\right)^{0.15} \left(\frac{1+z}{7}\right)^{0.75}, \tag{9.26}$$

which illustrates how rapidly accretion occurs at these very high redshifts. (In absolute terms, a $10^{10}\ M_\odot$ halo at $z = 7$ accretes gas at a rate of $2.6\ M_\odot\ {\rm yr}^{-1}$.)

Before discussing the fate of the cold flow gas in the small halos most important for high redshifts, we will briefly describe a long-standing problem for halos in the slow-cooling regime, $M > M_{\rm cool}$, that motivates much of the work on galaxy evolution. Although gas in this regime does indeed cool slowly, it is still relatively fast by cosmological standards. Thus, high-mass galaxies at low redshifts should still have accreted most of their baryons and formed stars from them. However, observations show an exponential decline in the number density of galaxies (with $L_\star$ comparable to the Milky Way luminosity) at mass scales well below the exponential cutoff in $n(m)$ (which occurs near galaxy cluster scales) at the present time. Evidently, then, some mechanism must prevent gas in massive halos from *overcooling* onto their central galaxies. This is likely to be feedback, either operating within the halos themselves or in the surrounding gas.

9.5.2 Halo Mergers

A second channel for adding mass to a galaxy is through a merger with a nearby halo. These merger rates are also described (roughly) by the extended Press-Schechter formalism, and equation (9.26) implicitly includes such growth in the overall accounting. However, the dynamics of such mergers differ greatly from smooth accretion.

Mergers are often divided into two classes, depending on the mass ratio of the merging systems. Consider m_1 and m_2 being the masses of the two systems, with $m_1 > m_2$. *Major mergers* are usually defined to have a mass ratio $m_1/m_2 < 4$. Such an interaction is quite dramatic, with so-called "violent relaxation" (due to time-varying gravitational potentials during the interaction) largely determining the structure of the resulting merger remnant, which may bear no resemblance to

the merging galaxies. (Indeed, the classical picture for the formation of elliptical galaxies is through major mergers of spirals.) *Minor mergers*, on the other hand, have $m_1/m_2 > 4$. The second system then makes only a small perturbation on the first, and the remnant retains the overall structure of the more massive object.

In particular, gravitational tides raised by mergers disrupt both the stars and gas inside the individual systems. The former can mix freely, but the latter collide via shocks, possibly triggering massive star formation over short timescales (typically a few dynamical times of the interacting galaxies). Such *starbursts* can be closely studied in nearby galaxies, and indeed often show evidence for strong gravitational perturbations. The rapid growth of high-redshift galaxies naively suggests that such starbursts may be very common in the early universe. However, mergers are difficult to model analytically, so they have most often been studied with numerical simulations. These show that equal mass mergers can roughly double the star formation efficiency over isolated systems when averaged over the merger time.

9.5.3 Disk Formation

As halo gas cools, it loses the pressure support holding it up against gravity and contracts to higher densities. This contraction continues until the gas reaches rotational support owing to its angular momentum.

The net angular momentum J of a galaxy halo of mass M, virial radius $r_{\rm vir}$, and total energy E, is commonly quantified in terms of the dimensionless spin parameter,

$$\lambda \equiv J|E|^{1/2}G^{-1}M^{-5/2}\,. \tag{9.27}$$

Expressing the halo rotation speed as $V_{\rm rot} \sim J/(Mr_{\rm vir})$ and approximating $|E| \sim MV_c^2$ with $V_c^2 \sim GM/r_{\rm vir}$, we find $\lambda \sim V_{\rm rot}/V_c$, i.e. λ is roughly the fraction of the maximal rotation speed above which the halo would break up.

After cooling the gas settles to a rotationally-supported disk. Let us write the disk mass as a fraction $\tilde{m}_d$ of the halo mass and let the disk angular momentum be a fraction $\tilde{j}_d$ of that of the halo. The scale radius of the disk is set by rotational support. As a simple estimate, let us take an isothermal profile for the dark matter halo and neglect the self-gravity of the disk. We further assume that the disk has an exponential surface density profile,

$$\Sigma(R) = \Sigma_0 \exp(-R/R_d), \tag{9.28}$$

with R_d the disk scale radius. The total disk mass is then $M_d = 2\pi\Sigma_0 R_d^2$. Because the circular velocity of an isothermal sphere is constant, the total angular momentum of the disk is

$$J_d = 2\pi \int V_c \Sigma(R) R^2 dR = 2M_d R_d V_c. \tag{9.29}$$

Setting this equal to a fraction $\tilde{j}_d$ of the total angular momentum of the halo as in equation (9.27), we obtain an expression for the disk scale length:

$$R_d = \frac{1}{\sqrt{2}} \left(\frac{\tilde{j}_d}{\tilde{m}_d} \right) \lambda\, r_{\rm vir}\,. \tag{9.30}$$

Note that the factor $\tilde{j}_d/\tilde{m}_d$ is simply the specific angular momentum of the disk material. The assumptions behind this simple expression are questionable: the self-gravity of the disk likely cannot be ignored once it collapses to a small size, the dark matter profile is not exactly isothermal (and it may respond to the gravity of the disk as well), and finally the disk may not have organized itself into a simple exponential profile. We also require some way to calibrate the specific angular momentum of the disk material and the spin parameter. The observed distribution of disk sizes in local galaxies suggests that the specific angular momentum of the disk is similar to that expected theoretically for dark matter halos, and so we assume $j_d/m_d = 1$. The distribution of disk sizes is then determined by the the distribution of spin parameters and halo masses. N-body simulations indicate that the former approximately follows a lognormal probability distribution,

$$p(\lambda)d\lambda = \frac{1}{\sigma_\lambda\sqrt{2\pi}} \exp\left[-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2}\right] \frac{d\lambda}{\lambda} \tag{9.31}$$

with $\bar{\lambda} = 0.05$ and $\sigma_\lambda = 0.5$.

Despite these difficulties, the simple model shows the expected scaling of the disk sizes with redshift: the size of a disk at a fixed halo mass is expected to scale as $R_d \propto (1+z)^{-1}$. Observations do indeed indicate that the luminous cores of galaxies follow this expected trend over the wide redshift range of $2 < z < 8$, as illustrated in Figure 9.14 (though note that these galaxies are binned by luminosity rather than mass).

For high-redshift galaxies, the primary lesson is that – even though the angular diameter distance *decreases* with z at high redshifts – the small masses and rapid cooling of the halo gas likely mean that the sources are extremely compact. Figure 9.15 shows the extrapolated relation between galaxy size and redshift, calibrated by current data on the size distribution and luminosity function of high-redshift galaxies. It implies that *JWST* will only be able to resolve galaxies at an AB magnitude limit $m_{\rm AB} < 31$ out to a redshift of $z \sim 14$. The next generation of large ground-based telescopes will resolve all galaxies discovered with *JWST*, but only if they are sufficiently clumpy to enable detection above the bright thermal sky.

9.5.4 Star Formation in Galaxies

Once the gas has cooled and collapsed to high densities, star formation can commence. Determining the conversion efficiency of gas to stars is the most important, and most challenging, aspect of galaxy formation. Nevertheless, theorists and observers have made enormous strides over the next decades in understanding the relevant processes, at least in the local Universe.

Traditionally, the star formation rate per unit area $\dot{\Sigma}_\star$ has been calibrated empirically as a function of the total gas surface density $\Sigma_{\rm gas}$. Observationally in the local Universe, these quantities correlate reasonably well over nearly seven orders of magnitude in surface density, with

$$\dot{\Sigma}_\star \propto \Sigma_{\rm gas}^n, \tag{9.32}$$

where $n \sim 1.4 \pm 0.1$. This so-called *Kennicutt-Schmidt relation* can also be interpreted in terms of a fixed fraction of the gas being converted into stars per orbital

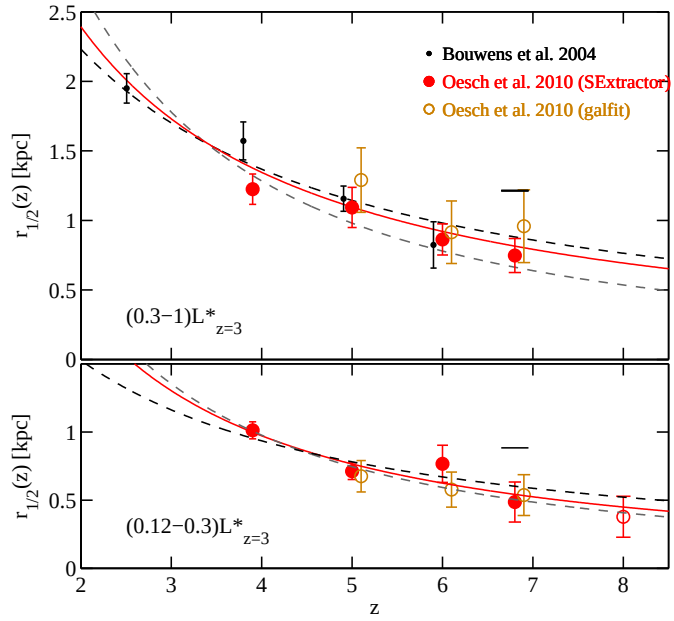

Figure 9.14 Observed evolution of the mean half-light radius of galaxies across the redshift range $2 < z < 8$ in two bins of fixed intrinsic luminosity: (0.3-1)$L_*(z = 3)$ (top) and (0.12-0.3)$L_*(z = 3)$ (bottom), where $L_*(z = 3)$ is the characteristic luminosity of a galaxy at $z = 3$ (Eq. 9.2). Different point types correspond to different methods of analysing the data. The dashed lines indicate the scaling expected for a fixed halo mass ($\propto (1 + z)^{-1}$; black) or at fixed halo circular velocity ($\propto (1+z)^{-3/2}$; gray). The central solid lines correspond to the best-fit to the observed evolution described by $\propto (1 + z)^{-m}$, with $m = 1.12 \pm 0.17$ for the brighter luminosity bin, and $m = 1.32 \pm 0.52$ at fainter luminosities. Figure credit: Oesch, P. A., et al. *Astrophys. J.* **709**, L21 (2010).

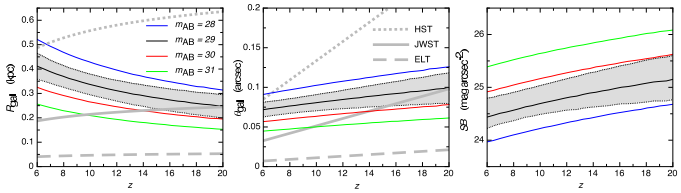

Figure 9.15 Theoretically extrapolated relation between galaxy size and redshift for four values of apparent AB magnitude. The *left* and *central* panels show the physical ($R_{\rm gal}$) and apparent angular sizes ($\theta_{\rm gal}$), respectively. The thick grey lines indicate the resolution of telescopes with diameters corresponding to *HST* (2.5 m), *JWST* (6.5 m) and a ground-based extremely large telescope or *ELT* (30 m).$The right panel shows the average surface brightness within a galaxy scale radius as a function of red$ 29 mag shows the 68% range of uncertainty on the mean. Figure credit: Wyithe, J. S. B., & Loeb, A. *Mon. Not. R. Astron. Soc.* **413**, L38 (2011).

time in the associated galactic disks. Despite the apparent success of this simple idea, as an empirical relation it must still be tested in other environments, and it is unclear whether star formation would obey the same relation at the low metallicity and low initial magnetization of the gas within the first galaxies.

Thus, a deeper understanding of star formation is highly desirable. As a first step, note that stars in the local Universe form in molecular clouds. One might therefore expect a more fundamental scaling of the star formation rate with the density of *molecular* (rather than atomic) gas. We write $f_{\rm H_2}$ for the fraction of molecular gas. Furthermore, local observations show that molecular clouds turn a constant fraction $\epsilon_{\rm ff} \approx 0.015$ of their gas into stars per free-fall time. This suggests a relation

$$\dot{\rho}_\star = \epsilon_{\rm ff} f_{\rm H_2} \rho / t_{\rm ff} \tag{9.33}$$

for the star formation rate. This relation then requires an estimate of the molecular fraction and the star formation efficiency parameter.

The derived relation is significantly more challenging than the analogous calculation in §5.1, because enriched gas has more channels for H_2 formation (particularly on the surface of dust grains), a much more complex radiation field (owing to the embedded star formation), dust shielding, and a turbulent, inhomogeneous ISM. The physical picture that emerges is one in which molecular gas is confined to the interiors of cold high-density gas complexes. We must then determine: *(1)* the relative mass of the cold phase, and *(2)* the fraction of the cloud able to go fully molecular. The latter is determined by balancing the rate of H_2 formation on dust grains with photodissociation by the (dust-extinguished) radiation field, similar to the calculations presented in chapter 5.

The fraction of gas in the cold phase is determined ultimately by the feedback from hot stars and supernovae. The canonical picture assumes a multiphase ISM, with a "hot" phase of diffuse ionized gas and a cold phase of dense star-forming gas (and likely an intermediate warm phase of atomic gas that can be ignored). Crudely, gas is exchanged between the phases (as well as the stellar component) through three basic processes: *(1)* star formation (from cold gas to stars), *(2)* cooling in the diffuse ISM (from hot to cold gas), and *(3)* supernovae (from stars and cold gas to hot gas). The last process includes not only the supernova ejecta itself but also cloud evaporation from conduction through the surrounding hot gas.

Radiative cooling is challenging to model in the ISM unless the galaxy is fully resolved, because the density (and possibly composition) is highly inhomogeneous. For example, simply assuming a uniform cooling rate (even enhanced by a clumping factor) throughout the entire galaxy would not allow *any* gas to cool to very low temperatures. In reality, cooling is highly inhomogeneous and subject to various thermal instabilities: because the cooling is most rapid in densest gas, this material will quickly cool and become neutral, while the low-density gas will remain hot. Fortunately, the simple assumption of a two-phase medium, each with a characteristic temperature, appears to provide a reasonable approximation to the full physics: in this case, the radiated energy determines the mass flow rate from the hot to cold phase.

Meanwhile, giant molecular clouds have much higher pressures than the surrounding ISM (at least in local galaxies), suggesting that their properties are set

by internal feedback processes rather than by coupling to the ISM. In particular, H II regions from embedded stars appear to provide the most important feedback mechanism. Because they are internally regulated, the properties of these clouds do not vary much between galaxies, which explains the apparent constancy of $\epsilon_{\rm ff}$ – though, of course, the conditions within high-redshift galaxies may be very different (in particular, the ambient ISM pressure will of course depend on its density).

Numerical simulations have shown results consistent with equation (9.32), except that $n \approx 2$ for massive galaxies and $n \approx 4$ for dwarfs, as required by recent data.

An alternative approach to this "bottom-up" view (which is fundamentally based on understanding the details of star formation) is to treat star formation within a global context. The basic idea is that star formation can only occur if some sort of global instability allows fragmentation to higher densities. The condition for this to occur is the *Toomre criterion* (see §5.2.3),

$$Q \equiv \frac{c_s \kappa_e}{\pi G \Sigma_g} < 1, \tag{9.34}$$

where $\kappa_e = (2\Omega/R)d(\Omega R^2)/dR$ is the epicyclic frequency for an angular frequency of rotation $\Omega(R) = v/R$ at a (cylindrical) radius R within the disk. However, once fragmentation begins, feedback from star formation will heat the gas, slowing further fragmentation. On the other hand, if star formation does not occur, the gas will cool rapidly, decreasing Q. The expectation (which appears to be realized in nearby galaxies) is therefore that galaxies will form stars sufficiently fast to maintain $Q \sim 1$. With a model for the feedback effects of stars, this provides an alternative method to determine $\dot{\Sigma}_\star$.

The two most obvious feedback mechanisms are radiation pressure from stars and supernovae, which "puff up" the disk and support the gas against the vertical component of gravity. Focusing on radiation pressure due to UV photons for concreteness, we can write (c.f. the momentum injection rate from radiation in §6.4.1)

$$p_{\rm rad} \sim (1 - f_{\rm esc})\epsilon \dot{\Sigma}_\star c, \tag{9.35}$$

where $\epsilon = 10^{-3}\epsilon_{-3}$ is the fraction of the baryonic rest energy converted to photons. (In nearby galaxies, supernovae produce a comparable pressure, but at high-redshifts the elevated ambient densities make them less important.)

Again let us take the simple model of an isothermal density profile within the halo with a 1D velocity dispersion σ, and assume that the disk contains a fraction f_g of the total matter. The fraction f_g is likely to be much larger than Ω_b/Ω_m, because the baryonic component has already cooled and collapsed into a disk. Assuming a thin disk, the vertical component of hydrostatic equilibrium can be written as $h \sim c_s/\Omega$, where Ω is the rotation rate. Writing $\Sigma_g = 2\rho h$, equation (9.34) provides an expression for the gas density inside the disk. With these two relations and $c_s^2 \sim p/\rho$, we can solve equation (9.35) for the required star formation rate to hold up the disk:

$$\dot{\Sigma}_\star \sim 10 \left(\frac{Q}{\epsilon_{-3}(1-f_{\rm esc})}\right)\left(\frac{f_g}{0.25}\right)^2 \left(\frac{\sigma}{70\ {\rm km\ s^{-1}}}\right)^4 \left(\frac{100\ {\rm pc}}{r}\right)^2 M_\odot\ {\rm yr^{-1}\,kpc^{-2}}, \tag{9.36}$$

where we have scaled σ to the appropriate circular velocity for a $10^{10}\ M_\odot$ galaxy at $z = 7$.

This particular estimate ignores the contribution from supernovae to the pressure and a possible enhancement in the radiation pressure from infrared emission by dust, but it gives a sense for how the global self-regulation criterion can be used to estimate the large-scale properties of galaxies. Such models typically connect more closely to the cosmological input parameters (the mass and accretion rate onto the dark matter halo). For example, given $\dot{M}$ one can integrate the star formation rate inwards and determine the gas fraction f_g at each radius self-consistently. The advantage of this approach is that it does not require a calibration to local galaxies and so is more robust to any unknown changes in the small-scale physics of star formation at high redshift; the disadvantage is that it makes strong assumptions about Q, the structure of the disk, and the relation between star formation feedback and the disk properties (ignoring other sources of pressure support like turbulence, for example).

However, the above relation does produce a surface-density law consistent with local models. Defining $\epsilon_{\rm SFR}$ via $\dot{\Sigma}_\star = \epsilon_{\rm SFR}\Sigma_g\Omega$, where Ω is the angular velocity (comparable to the dynamical time, and roughly the growth rate of global instabilities in disk galaxies), self-regulation at $Q \sim 1$ via radiation pressure yields $\epsilon_{\rm SFR}$ of a few percent for moderately large galaxies, with a predicted scaling $\epsilon_{\rm SF} \propto \Sigma_g$.

9.5.5 Black Hole Growth in Galaxies

As described in chapter 7, it is now well-established that nearly all present-day galaxies with spheroids also have supermassive black holes in their centers. Because the properties of these black holes correlate with their host galaxies, it is natural to include them in models of galaxy formation and evolution. (We will also see below that they may be important sources of feedback in some galaxies.)

Black holes may be fed smoothly and relatively slowly during the normal growth of a galaxy: some small fraction of the accreted gas may sink all the way through the galaxy and be swept into the black hole. The *minimal* accretion rate is given by the Bondi estimate from §5.2.1, $\dot{M}_{\rm BH} \sim G^2 M_{\rm BH}^2 \rho / c_s^3$, where the density ρ and sound speed c_s are evaluated at the accretion radius $R_{\rm acc} \sim G M_{\rm BH}/c_s^2$. However, this generally produces slow accretion.

A more efficient method of feeding black holes is through a mechanism that channels gas toward the black holes. This can include any global instability (such as bars or spiral waves), but in the cosmological context galaxy mergers are often identified as the most likely mechanism. As described previously, the torques generated during such mergers can be large, and a fair fraction of this gas can be fed toward the center of the remnant according to numerical simulations. Dimensionally, such strong torques will produce an inward radial velocity that is some fraction of the local sound speed (~ 0.2 for spiral waves).

However, the fate of this gas is difficult to determine analytically, because it is of course also subject to star formation and feedback. Global disk models (as described in the previous section) can in principle follow the gas toward the galaxy's center, but a more phenomenological approach is often taken by assuming that

the $M_{\rm BH}$–σ relation holds for all spheroidal galaxies and using it to *assign* a total accreted mass following a merger. As usual, one must worry about whether this relation holds during the earliest phases of galaxy formation (and in particular how it extrapolates with redshift).

Of course, if *both* galaxies in a merger have black holes, the resulting system will likely host a binary black hole. The fate of this merging system is described in §7.4, and it may have interesting signatures even beyond electromagnetic radiation. If the binary does not coalesce before the next merger, a triple (or higher multiple) system would form, from which the lightest black hole may be ejected at a speed of thousands of $\mathrm{km\ s^{-1}}$. Due to the increase in merger rate at high redshifts, multiple black hole systems are expected be more common in early galaxies.

9.5.6 Feedback and Galaxy Evolution

Feedback from stars and black holes is crucial for galaxy evolution models in at least three respects. First, as we have already seen, it is crucial in setting the properties of the star-forming gas within the galaxy itself, through radiation from stars and mechanical energy input via supernovae. Second, it enriches the gas, changing its dust content, cooling rates, and stellar properties. Finally, winds (whether driven by radiation pressure or supernovae) offer another end point (other than stars) for accreted gas: it can be ejected entirely from the halo.

Modeling these different aspects is clearly very challenging, and often it is parameterized in simpler fashions. For example, we have already seen that the "internal" feedback regulating star formation can be implicitly included in star formation laws with relatively simple phenomenological prescriptions like the Kennicutt-Schmidt law or its more recent modifications (though, again, one must always worry whether such prescriptions can be extrapolated robustly to the high redshifts of interest to us).

Chemical enrichment is the most straightforward of these effects to model: given an initial mass function, the rate at which material is enriched and returned to the ISM is straightforward to calculate. The ejecta are typically assumed to mix efficiently with the ISM, so that future generations of stars have monotonically increasing metallicity. The major uncertainties in chemical evolution are the properties of the gas accreting onto the galaxy (whether it is pre-enriched) and the fraction of the ejected metals entrained into winds and carried out of the galaxy.

Perhaps the most significant aspect of feedback is mass loss through winds, which can dramatically affect the overall star formation efficiency in small galaxies. We have already discussed the complex physics of winds in §6.4. We expect feedback to be most important in the small gravitational potential wells of low-mass galaxies. The most crucial question is how the wind efficiency varies with galaxy mass, which depends largely on the underlying physics (i.e., momentum-driven or energy-driven). For example, suppose that the supernova energy accelerates a fraction of ISM material (at a rate $\dot{M}_w$) to the escape speed of the dark matter halo. Then we have $\dot{M}_w \propto \dot{M}_\star \omega_{\rm SN}/v_{\rm esc}^2$. If, on the other hand, much of the energy is lost through cooling so that the momentum input of the supernovae or radiation pressure from the stars drive the wind, then $\dot{M}_w \propto \dot{M}_\star/v_{\rm esc}$, a significantly gentler

dependence on $v_{\rm esc} \propto M^{1/3}$.

We note that beyond this overall scaling, the mass loading of the winds is also highly uncertain, because the total matter content of winds is difficult to observe. So far, the best evidence comes from redshifted metal lines in galaxy spectra, which at best provide a velocity and column density of the material; without the distance of the absorbing material from the galaxy, the total mass is difficult to assess. Typically, however, the mass loss rate is assumed to be roughly equal to the star formation rate.

Finally, winds not only entrain gas but can also prevent circumgalactic gas from accreting onto the galaxy by heating it. This reduces the inflow rate onto the galaxy.

Given the basic energetics of the process, the prevailing expectation is that supernova feedback may suppress star formation in small galaxies and help explain the relative dearth of low-luminosity galaxies compared to the number of small-mass halos. This is consistent with local observations, where the total stellar mass is $\propto M^{2/3}$ for $M < M_{\rm crit} \sim 3 \times 10^{10}\ M_\odot$ and constant above it. Assuming that $E_{\rm SN} \propto M_\star$, and that star formation continues until supernovae clear the halo of its remaining baryons by injecting an energy comparable to the binding energy of the gas, we would expect $M_\star/M \propto V_c^2 \propto M^{2/3}$. If this explanation applies at lower redshifts as well, we would expect a similar suppression there, in galaxies with $V_c < 100$ km s^{-1}.

We should also consider feedback from black holes during accretion episodes of quasar activity. This is primarily important in driving galactic winds. As we have described in §**??**, the energy input from quasars can exceed that from star formation, although the coupling of this energy to the ISM is not yet understood. (In some cases, such as jets from radio quasars, the energy may escape the galaxy in narrow channels without clearing all the gas.) On the other hand, the tight $M_{\rm BH}$–σ relation is highly suggestive of a fundamental relationship between the growth of black holes and their host galaxy's stars. Because (at least naively) this relationship suggests that the black hole mass scales superlinearly with halo mass, this feedback channel is *more* effective in larger galaxies and is often invoked as a potential solution to the "overcooling" problem in massive, low-redshift galaxies.

Given the common "merger models" for AGN growth, one plausible physical picture is that the merger funnels large quantities of gas toward the remnant's center, triggering a starburst. Some of the gas continues to fall inward and is accreted by the black hole, which drives a wind outward into the galaxy. Once the black hole grows large enough, this wind unbinds the remaining gas and halts the star formation episode, at least until another major accretion event occurs. This scenario naturally explains many aspects of the low-redshift Universe (such as the relation of black holes to spheroids rather than disks), but its application to very high redshifts – where spheroids may or may not even exist, and the much more rapid growth of galaxy-sized halos likely prevents active black holes from entirely halting star formation – is far from clear.

9.5.7 From Galaxy Model to Stellar Spectra

In addition to the raw star formation rate, most observables depend on the initial mass function (IMF) of the stars. We have already discussed this in SS5.2.4 and 5.3.3, where we described how local measurements are consistent with a (broken) power law in the stellar mass range of ~ 0.1–$100\ M_\odot$. Once true galaxies form, with reasonably enriched gas, the IMF likely approaches this form, though (as we argued before) the characteristic mass may increase at higher redshifts owing to the higher CMB temperature.

An additional issue that appears to be important for generating realistic stellar populations is the finite mass of the gas clouds from which stars form, as it now appears that most stars form in groups (though they may later disperse). The range of allowable star cluster masses is called the *cluster initial mass function*; local observations show are consistent with a power-law distribution of number per unit mass of slope ~ -2 (reflecting an equal amount of mass per logarithmic mass bin) between a few tens to $\sim 10^6\ M_\odot$. This is important because the total fuel mass may limit the maximum stellar mass that can form in that environment – in other words, even if the underlying stellar IMF stretches smoothly to very high masses, a dearth of high-mass gas clouds will cause a dearth of high-mass stars. More accurate stellar population models can be stochastically sampling the cluster initial mass function to generate a set of star clusters and then stochastically sampling the stellar IMF (taking into account the maximum stellar mass allowed within each one).[v]

Given a metallicity and following this procedure for a set of stars forming at a particular instant, one can then calculate how the luminosity and spectrum of the population evolve with time using libraries of stellar models. Although isolated, non-rotating stars are well understood, there remain some important uncertainties in this modeling. The fundamental challenge is that the ionizing luminosity comes from only a small fraction of the stars (those with the highest masses). Thus, small variations in their formation efficiency or properties can cause substantial uncertainties in the models.

For example, the ionizing flux of stars cannot be observed directly and instead must be modeled from their feedback effects on surrounding H II regions. Meanwhile, the massive stars responsible for these photons have atmospheres that are out of local thermodynamic equilibrium and often undergo substantial mass loss through line-driven winds. These so-called *Wolf-Rayet stars* present particular challenges to models.

As a second example, most ($> 75\%$) stars are born with neighbors (as binaries or even larger multiple systems) in the local Universe. Binarity can dramatically affect the evolution of the component stars. For example, suppose one (the more massive star) reaches its supergiant phase first. It expands rapidly, with some of its envelope

[v] It might seem more natural to a cosmologist to define the IMF as the net result of this process, since that is the galaxy-wide initial mass function of stars. However, in order to measure the stellar IMF one must find a population of stars that formed simultaneously – in other words, a single cluster. Thus, the canonical stellar IMF – measured long before the importance of the cluster IMF was recognized – is only part of the "real" mass distribution of stars.

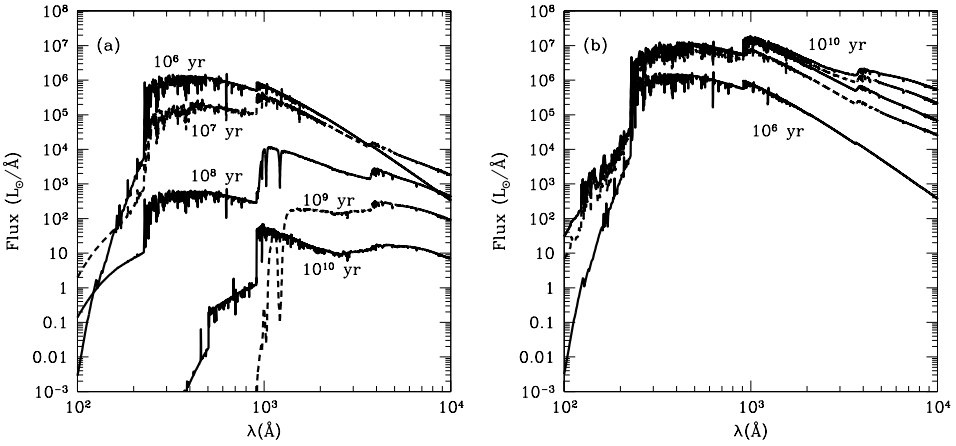

Figure 9.16 Spectral synthesis models of stellar populations. *(a)* Spectra for an instantaneous burst of star formation with $M_\star = 10^6\ M_\odot$. *(b)* Spectra for a constant star formation rate with $\dot{M}_\star = 1\ M_\odot\ \mathrm{yr}^{-1}$. Both panels show predicted spectra for populations 10^6, 10^7 10^8, 10^9, and 10^{10} yr after the onset of star formation. The calculation assumes $Z = 0.05\ Z_\odot$, includes binaries, and ignores nebular reprocessing in all cases. It adopts an IMF with a slope of -1.3 for 0.1–0.5 $M_\odot$ and -2.35 for 0.5–120 $M_\odot$, and does not account for the finite mass of star-forming clouds. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

passing the Roche limit and escaping. The "naked" surface of this more massive star will then become hotter, producing more ionizing photons. Meanwhile, the neighbor may accrete some of this additional mass and itself become more massive (and hence hotter) and possibly gain angular momentum and rotate faster (which also tends to make it hotter).

Overall, stellar models vary by a factor of a few in their ionizing flux, even at a fixed metallicity and stellar IMF. They are generally more consistent with each other at longer wavelengths, but the differences can still be important. Nevertheless, several general trends are apparent:

- **Stellar age:** Because the most massive stars have the shortest lifetimes, the spectrum (particularly at high frequencies) is extremely sensitive to the elapsed time since a star formation episode. Figure 9.16*a* shows this explicitly. After only 1 Myr, many stars have not yet evolved into their hot phases, and so the ionizing flux is relatively small. The ionizing spectra harden shortly afterward and then rapidly fade away as the hot stars die. Meanwhile, the continua also fade steadily as more stars explode in supernova.

- **Star formation history:** A corollary of the previous point is that spectral measurements can determine the star formation history of a galaxy. There

is, however, the important possibility that star formation may not be instantaneous. If it instead continues at a constant rate for a long time period (i.e., much longer than the age of the most massive stars), the high-energy photons will still be sensitive to the high-mass, short-lived stars (and hence the current star formation rate), but the lower-energy photons will depend on the integrated population of low-mass stars and so measure the total stellar mass. Figure 9.16*b* shows spectra with ongoing star formation over timescales of 10^6, 10^7, 10^8, 10^9, and 10^{10} years (from lower to upper curves). Note how the spectra roughly converge after long times, only increasing at very long wavelengths as the galaxy continues to accumulate more and more low-mass stars.

However, we should emphasize that the starburst and constant star formation rate histories are only simple examples, and of course more detailed observations can constrain more complex histories. For the high redshifts of interest to us, where galaxies grow extremely rapidly, the so-called "exponential" histories, where SFR$\propto e^{-t/t_\star}$, may also be appropriate.

- **Metallicity:** In general, the higher opacities of heavy elements lead to slightly cooler stellar atmospheres and hence redder spectra. Of course, they also change the spectral lines substantially. Figure 9.17 illustrates this for low and high metallicity models (see also §5.4 for a comparison to Population III models). Although the long wavelength tail is nearly unchanged, increasing the metallicity decreases the ionizing flux by a factor of up to several. The non-trivial differences amongst these spectra indicate that the metallicity is an observable quantity given high-resolution spectra. However, one must bear in mind that metallicity is likely to evolve as star formation proceeds, since it is the stars themselves that enrich the medium.

- **Binaries:** Finally, we have already mentioned that the inclusion of binary evolution can substantially modify the far-ultraviolet fluxes of stellar populations. Figure 9.17*b* shows this explicitly. Binaries change only slightly the long-wavelength flux but increase the ionizing flux significantly.

9.6 OBSERVATIONAL SIGNATURES OF THE ISM

In the previous section we saw how "synthesized" galaxy spectra can be created given information about the star formation history and properties of the stars themselves. Of course, the other major component of the galaxy – its ISM – also has important observable consequences that can affect both the observed continuum of the stars and, especially, the galaxy's spectral lines. A firm understanding of these effects is necessary to understand the stellar component, but it also allows us to learn about the diffuse component of the galaxy and hence its fuel supply and feedback processes. Here, we will briefly outline the most important of these effects.

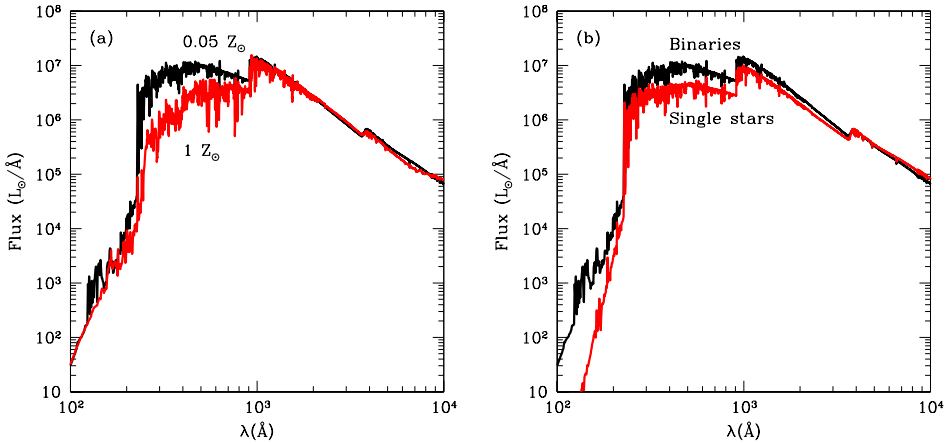

Figure 9.17 Spectral synthesis models of stellar populations: *(a)* variation with metallicity; *(b)* contribution of stellar binaries (assuming $Z = 0.05\ Z_{\odot}$). Both panels show predicted spectra for a constant star formation rate of $1\ M_{\odot}\ \mathrm{yr}^{-1}$, 10^8 yr after the star formation began. The calculation ignores nebular reprocessing in all cases, adopts an IMF with a slope of -1.3 for the stellar mass range of 0.1–0.5 $M_{\odot}$ and a slope of -2.35 for masses between 0.5–120 $M_{\odot}$. It does not account for the finite mass of star-forming clouds. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

9.6.1 Nebular Emission Lines

The raw stellar spectra computed in §9.5.7 likely do not reach an observer without substantial changes from their surroundings. The most immediate is the interaction of ionizing photons with the local ISM: presuming that the stars form in dense environments, many of those photons will be absorbed by local hydrogen or helium atoms. In chapter 8, we parameterized the fraction that escape their host galaxy with $f_{\rm esc}$ and saw that this is at most a few percent in low redshift galaxies. The remaining photons ionize atoms in their host galaxy, which then undergo radiative cascades, reprocessing the energy originally contained in ionizing photons into emission lines at longer wavelengths.

Figure 9.18 shows two examples of this reprocessing, for two different assumed metallicities (note that the solar metallicity curve has been shifted down by a factor of 100 for clarity of presentation; its continuum amplitude is in reality just slightly smaller). The strengths of these recombination lines are determined by ionization balance in the H II regions. Assuming that they are Strömgren spheres, the total number of recombinations per second is determined by the total number of ionizations and so it measures the ionizing luminosity. The relative strengths of the hydrogen (and helium lines, for very low metallicities and hot stars) depend on atomic physics and so provide a measure of the temperature of the gas.

However, metal lines can also be important diagnostics, if they do exist. These are usually collisionally excited forbidden transitions, such as [O II], [O III], and [N II]; they are important because such ions have excitation temperatures $\sim 10^4$ K, comparable to the expected temperatures of stellar H II regions. The ratios of the strength of these emission lines to those of hydrogen depend on the (gas) metallicity and can be used to estimate it; this has proven to be very useful at lower redshifts, though it is not yet possible at $z > 4$.

Nebular emission lines also offer a useful probe of the escape fraction, because their strength is proportional to $(1 - f_{\rm esc})$. Reprocessing shifts photons from the short wavelength tail to longer wavelengths and so can even change broadband colors (i.e., a significant fraction of the energy measured in a particular observational filter may actually be contained in emission lines rather than the raw stellar continuum). For example, suppose one estimates a spectral index $f_\lambda \propto \lambda^{-\beta}$ from the average broadband colors. The difference between full nebular reprocessing and full escape corresponds to a range in $\beta \sim 2.2$–3.1 for very young stellar populations (< 3 Myr), though the difference falls to ~ 0.1 for older populations (> 100 Myr). However, even when $f_{\rm esc} \sim 1$, photons with $\lambda < 912$ Åwill not be directly observable, because they will quickly be absorbed by the intervening IGM at $z > 5$.

9.6.2 Dust

The most obvious effect on the stellar spectrum comes from dust, which absorbs stellar photons (especially those with short wavelengths), heats up, and ultimately re-radiates that energy in the infrared or submillimeter bands. The effects of dust depend on its total mass, its composition, and the relative geometry of the stel-

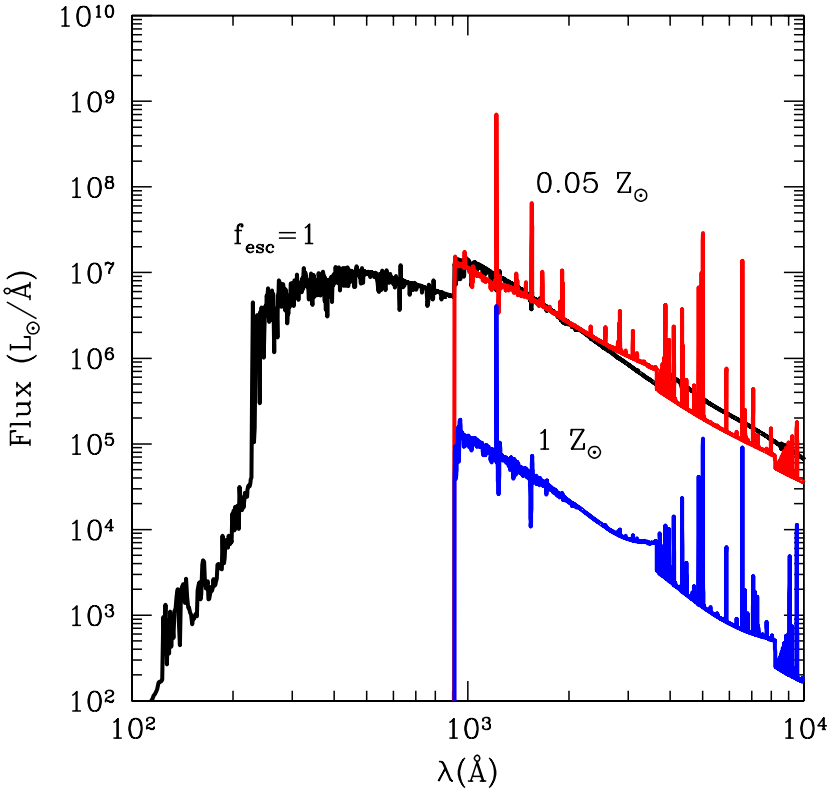

Figure 9.18 Effects of host galaxy absorption on stellar spectra. The curve extending to short wavelengths is the raw stellar spectrum (identical to the curve with binaries in Fig. 9.17*b*). The other two curves show the spectrum assuming that all ionizing photons are absorbed by the galaxy ISM and reprocessed into emission lines at longer wavelengths. The upper curve assumes $Z = 0.05\ Z_{\odot}$; the lower curve assumes solar metallicity for both the stars and ISM. The latter is shifted down by two orders of magnitude for clarity of presentation. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

lar and dust components of the ISM. The total dust mass determines the overall extinction of the gas, while the composition of the dust determines the relative extinction across different wavelengths. Unfortunately, this so-called *extinction law* is found to vary even amongst nearby galaxies, particularly for short wavelengths. Given that high-redshift galaxies are much earlier in their star formation history, one would also expect that their dust may have very different compositions from those in the present-day Universe. Moreover, if the dust preferentially surrounds star-forming regions it will have a larger effect on young, hot stars than on the low-mass stars that may have wandered far from their birth sites. Thus, predicting the dust absorption from early galaxies is rather difficult.

The dust emission is equally interesting. The dust will radiate thermally, though the spectrum will not typically be a true blackbody, because dust in different environments may have different temperatures (locally, ranging from $\sim$ 20–40 K in the low-density IGM up to several hundred K in star-forming regions) and because of the range of dust particle sizes (the blackbody approximation is not valid for wavelengths smaller than the particle radius). In a simple model, the dust emission spectrum can be parameterized by two quantities: *(1)* the dust temperature T_d, and *(2)* the dust emissivity $\epsilon_{\nu,\rm dust}$.

The dust temperature T_d is set by balancing the incident energy with the dust emission. In the simplest model, we assume blackbody emission and write,

$$T_d^4 \approx T_{\rm CMB}^4 + T_m^4 + T_\star^4 + T_{\rm AGN}^4, \tag{9.37}$$

where the four terms account for the CMB radiation field, non-radiative energy input (via cosmic rays or supernovae), the stellar radiation field, and any energy input from AGN (which appears to be important in some galaxies). The last two quantities presumably scale with the surface density of star formation and black hole accretion rate, respectively, though locally they appear to saturate at ~ 60 K and ~ 150 K, respectively. The CMB contribution is rarely important at low redshifts, but it will become much more significant at higher redshifts.

The dust emissivity is often approximated as a power law, $\epsilon_{\rm dust}(\nu) \propto \nu^\beta$, with $\beta \sim 1$ at high photon frequencies ν (in order to match observations) and $\beta \to 2$ at long wavelengths from standard scattering theory. If the dust is optically thin, the spectrum will follow $f_{\nu,\rm dust} \propto \epsilon_{\nu,\rm dust} B_\nu$, and the normalization will be determined by balancing the input luminosity (from stars or AGN) with this thermal emission. At low and moderate redshifts, some very rapidly star-forming galaxies have such high dust content that nearly all of their emission comes in the infrared and sub-millimeter bands. Whether more distant analogs for these exist is so far unknown and depends primarily on how quickly galaxies can build large dust columns.

Although it is clearly difficult to predict from first principles, this dust emission has one very important property from an observer's perspective: the spectra of dusty star-forming galaxies are such that, in the sub-millimeter band, the observed fluxes will be nearly independent of redshift well into the cosmic dawn. This occurs because the peak of the blackbody spectrum usually lies blueward of the observational bands, so it moves into the observed bands as the galaxy's redshift increases. Such a *negative K-correction* makes sub-millimeter observations potentially extremely powerful for observing distant galaxies. Figure 9.19 illustrates this for a

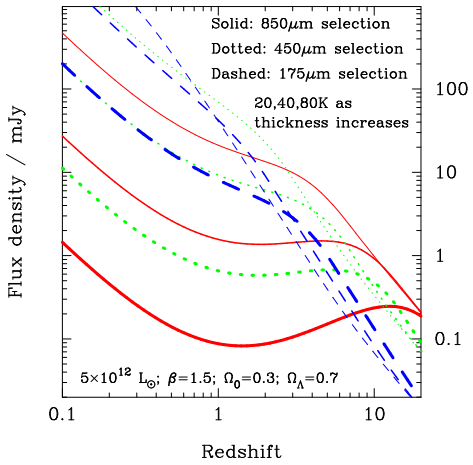

Figure 9.19 Observed flux density as a function of redshift in three sub-millimeter bands, for several different fiducial dust temperatures. The solid, dotted, and dashed lines assume observations in the 850, 450, and 175 μm bands, respectively. The three curves within each set take different dust temperatures, $T = 20$, 40, and 80 K, from thick to thin lines. All assume dust emissivity with $\beta = 1.5$. Figure credit: Blain, A. W. et al., *Phys. Rep.* **369**, 111 (2002).

model galaxy based on a local composite of dust-dominated galaxies. It shows how the observed flux for galaxies in three different bands, and taking three different fiducial dust temperatures, varies with redshift. Interestingly, at the longest wavelengths and/or lowest dust temperatures, the flux hardly varies with redshift: if a telescope (such as ALMA) can detect a given galaxy population at $z \sim 1$, it may be able to detect similar galaxies all the way to $z \sim 10$.

9.6.3 Interstellar Absorption Lines

In addition to metal emission lines from H II regions (see §9.6.1), metals in the ISM will also cause absorption lines in a galaxy spectrum. In principle, these are interesting for measuring the gas-phase metallicity of the ISM; however, in lower redshift galaxies the strongest lines tend to be saturated (as measured by the relative strengths of doublet lines), which makes such a measurement extremely difficult.

Instead, these absorption lines are useful for measuring the properties of galactic winds. Interestingly, although many of the lines appear saturated, they do *not* completely attenuate the starlight. The depth of the absorption therefore tells us the covering fraction of the high-metal-column gas. Meanwhile, these absorption lines are nearly always redshifted, as would be expected for gas flowing out of the galaxy along the line of sight toward the observer. These lines (together with Lyman-α, which we will discuss in chapter 10) provide the best direct evidence for galactic outflows at lower redshifts. However, their interpretation remains extremely difficult because they provide no information on how far the gas has traveled from the galaxy.

9.6.4 Radio Emission Lines

Another important tracer of the gas phase is emission from molecular and atomic lines: these provide a significant fraction of the cooling radiation that escapes galaxies, especially in star forming regions. We will consider two important examples here: CO, which is an excellent tracer of star formation in the local Universe (and at moderate redshifts), and the [C II] fine structure line (with a rest wavelength of 157.7 μm), which contains ~ 0.1–1% of the bolometric luminosity of nearby star-forming galaxies. Table 9.1 lists many other possible transitions, together with their approximate (local) relation between luminosity and star formation rate.

CO has a ladder of rotational levels $J \rightarrow (J-1)$ with frequencies $\nu_J = J\nu_{\rm CO}$, where $\nu_{\rm CO} = 115.3$ GHz, which corresponds to an excitation temperature of a $T_{\rm CO} = 5.5\,{\rm K}$. This low temperature means that CO is excited even in the cold, dense molecular clouds out of which stars form. Moreover, because carbon and oxygen are relatively abundant, it is by far the strongest metal line in such regions. At low to moderate redshifts, there is a tight correlation between CO luminosity (here expressed in the 1-0 transition) and the star formation rate,

$$L_{\rm CO(1-0)} = 3.2 \times 10^4 L_\odot \left(\frac{\rm SFR}{M_\odot\ {\rm yr}^{-1}} \right)^{3/5}. \tag{9.38}$$

As usual, it is not clear if this relation can safely be extrapolated to high redshifts.

To predict the CO luminosity on more physically motivated grounds, we need to know the molecule's abundance as well as its excitation temperature. The latter is set by the cloud's dust (T_d in equation 9.37). The abundance may not be as important as it seems: in local galaxies, giant molecular clouds are optically thick in CO, so decreasing the CO content does not (at first) decrease the overall luminosity. However, note that the dust temperature in equation (9.37) does implicitly depend on the metal content, because once a cloud becomes optically thin to the stellar photons the dust temperature decreases. This will in turn decrease the CO luminosity. In fact, nearby low-metallicity galaxies fall well below the relation in equation (9.38), though the much more compact high-redshift galaxies may have very different characteristics.

Moreover, because $T_{\rm CO}$ is so small, many individual levels could be excited and so many transitions could be visible. In local thermodynamic equilibrium, the line ratios just depend on temperature, but at different temperatures and densities the higher levels may not be thermalized. Ideally, one would then like to observe a wide range of lines in order to fully characterize the molecular clouds.

An alternative bright probe is the fine-structure 157.7 μm line of [C II], which is much less sensitive to the chemistry of the molecular clouds. This line, which arises from a ${}^2P_{3/2} \rightarrow {}^2P_{1/2}$ electronic transition, has an excitation temperature set primarily by collisions with free electrons and interactions with CMB photons, so it can be predicted much more robustly: the primary uncertainty is simply the mass of atomic carbon, or the metallicity of the gas. For $z > 6$, the [C II] line is redshifted into the sub-millimeter or millimeter range and may be observed with the ALMA observatory.

Table 9.1 Prominent interstellar emission lines in star forming galaxies, along with their typical ratio R between the luminosity and star formation rate (in units of $L_\odot/(M_\odot/\mathrm{yr})$). For the first 7 lines R is measured from a sample of low redshift galaxies; the other lines have been calibrated based on the galaxy M82. Table credit: E. Visbal & A. Loeb, *JCAP* **11**, 16 (2010).

Species	Emission Wavelength[μm]	R[$L_\odot/(M_\odot\ \mathrm{yr}^{-1})$]
CII	158	6.0×10^6
OI	145	3.3×10^5
NII	122	7.9×10^5
OIII	88	2.3×10^6
OI	63	3.8×10^6
NIII	57	2.4×10^6
OIII	52	3.0×10^6
^{12}CO(1-0)	2610	3.7×10^3
^{12}CO(2-1)	1300	2.8×10^4
^{12}CO(3-2)	866	7.0×10^4
^{12}CO(4-3)	651	9.7×10^4
^{12}CO(5-4)	521	9.6×10^4
^{12}CO(6-5)	434	9.5×10^4
^{12}CO(7-6)	372	8.9×10^4
^{12}CO(8-7)	325	7.7×10^4
^{12}CO(9-8)	289	6.9×10^4
^{12}CO(10-9)	260	5.3×10^4
^{12}CO(11-10)	237	3.8×10^4
^{12}CO(12-11)	217	2.6×10^4
^{12}CO(13-12)	200	1.4×10^4
CI	610	1.4×10^4
CI	371	4.8×10^4
NII	205	2.5×10^5
^{13}CO(5-4)	544	3900
^{13}CO(7-6)	389	3200
^{13}CO(8-7)	340	2700
HCN(6-5)	564	2100

9.7 GRAVITATIONAL LENSING

Another approach adopted by observers benefits from magnifying devices provided for free by nature, so-called "gravitational lenses." Rich clusters of galaxies have such a large concentration of mass that their gravity bends the light-rays from any source behind them and magnifies its image. This allows observers to probe fainter galaxies at higher redshifts than ever probed before. The redshift record from this method is currently associated[38] with a strongly lensed galaxy at $z = 7.6$. As of the writing of this book, this method has provided candidate galaxies with possible redshifts up to $z \sim 10$, but without further spectroscopic confirmation that would make these detections robust.[39]

The chance alignment of a foreground object along the line of sight to a high redshift source could result in the magnification, distortion, and potentially splitting of the source image due the deflection of its light rays by the gravitational field of the foreground object. The probability for *gravitational lensing* grows with increasing source redshift, due to the increase in the path length of the source photons. Although the lensing probability is only of anecdotal significance of $< 1\%$ for sources at $z < 2$, its magnitude could rise by an order of magnitude and affect the statistics of bright sources during the epoch of reionization.

Assuming that the gravitational potential of the lens is non-relativistic $|\Phi/c^2| \ll 1$, the effect of spacetime curvature on the propagation of light rays is equivalent to the effect of an effective index of refraction n,

$$n = 1 - \frac{2}{c^2}\,\Phi. \tag{9.39}$$

This follows from the deviation imparted to the phase of the electromagnetic wave by the potential of the lens (relative to a flat spacetime metric). The lens potential Φ is negative and approaches zero at infinity. As in normal geometrical optics, a refractive index $n > 1$ implies that light travels slower than in vacuum. Thus, the effective speed of a ray of light in a gravitational field is

$$v = \frac{c}{n} \simeq c - \frac{2}{c}\,|\Phi|\ . \tag{9.40}$$

The total time delay Δt, so-called the *Shapiro delay*, is obtained by integrating over the light path from the observer to the source:

$$\Delta t = \int_{\rm source}^{\rm observer} \frac{2}{c^3}\,|\Phi|\,dl\ . \tag{9.41}$$

A light ray is defined as the normal to the phase front. Since Φ and hence the phase delay of the electromagnetic wave vary across the lens, a light ray will be deflected by the lens as in a prism. The deflection is the integral along the light path of the gradient of n perpendicular to the light path, i.e.

$$\vec{\hat{\alpha}} = -\int \vec{\nabla}_\perp n\,dl = \frac{2}{c^2}\int \vec{\nabla}_\perp \Phi\,dl\ . \tag{9.42}$$

In all cases of interest the deflection angle is very small. We can therefore simplify the computation of the deflection angle considerably if we integrate $\vec{\nabla}_\perp n$ not along

the deflected ray, but along an unperturbed light ray with the same impact parameter (with multiple lenses, one takes the unperturbed ray from the source as the reference trajectory for calculating the deflection by the first lens, the deflected ray from the first lens as the reference unperturbed ray for calculating the deflection by the second lens, and so on).

The simplest lens is a point mass, M, with a Newtonian potential,

$$\Phi(b,z) = -\frac{GM}{(b^2+z^2)^{1/2}} \ , \tag{9.43}$$

where b is the impact parameter of the unperturbed light ray, and z indicates distance along the unperturbed light ray from the point of closest approach. We therefore have

$$\vec{\nabla}_\perp \Phi(b,z) = \frac{GM\,\vec{b}}{(b^2+z^2)^{3/2}} \ , \tag{9.44}$$

where $\vec{b}$ is orthogonal to the unperturbed ray and points toward the point mass. Equation (9.44) then yields the deflection angle

$$\hat{\alpha} = \frac{2}{c^2}\int \vec{\nabla}_\perp \Phi \, dz = \frac{4GM}{c^2 b} \ . \tag{9.45}$$

Since the Schwarzschild radius is $R_{\rm Sch} = (2GM/c^2)$, the deflection angle is simply twice the inverse of the impact parameter in units of the Schwarzschild radius. As an example, the Schwarzschild radius of the Sun is 2.95 km, and the solar radius is 6.96×10^5 km. A light ray grazing the limb of the Sun is therefore deflected by an angle 8.4×10^{-6} radians $= 1.''7$.

The deflection angle from more a complicated mass distribution can be treated as the sum over the deflection caused by the infinitesimal point mass elements that make the lens. Since the deflection occurs on a scale $\sim b$ which is typically much shorter than the distances between the observer and the lens or the lens and the source, the lens can be regarded as thin. The mass distribution of the lens can then be replaced by a mass sheet orthogonal to the line-of-sight, with a surface mass density

$$\Sigma(\vec{\xi}) = \int \rho(\vec{\xi},z)\, dz \ , \tag{9.46}$$

where $\vec{\xi}$ is a two-dimensional vector in the lens plane. The deflection angle at position $\vec{\xi}$ is the sum of the deflections from all the mass elements in the plane:

$$\vec{\hat{\alpha}}(\vec{\xi}) = \frac{4G}{c^2}\int \frac{(\vec{\xi}-\vec{\xi}')\Sigma(\vec{\xi}')}{|\vec{\xi}-\vec{\xi}'|^2}\, d^2\xi' \ . \tag{9.47}$$

In general, the deflection angle is a two-component vector. In the special case of a circularly symmetric lens, the deflection angle points toward the center of symmetry and has an amplitude,

$$\hat{\alpha}(\xi) = \frac{4GM(\xi)}{c^2\xi} \ , \tag{9.48}$$

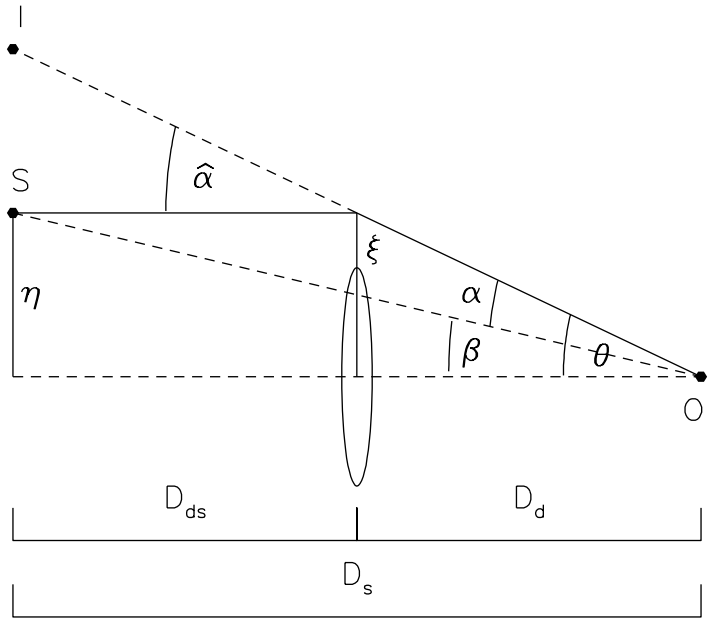

Figure 9.20 Geometry of gravitational lensing. The light ray propagates from the source S at transverse distance η from an arbitrary axis to the observer O, passing the lens at transverse distance ξ. It is deflected by an angle $\hat{\alpha}$. The angular separations of the source and the image from the axis as seen by the observer are β and θ, respectively. The distances between the observer and the source, the observer and the lens, and the lens and the source are $D_{\rm s}$, $D_{\rm d}$, and $D_{\rm ds}$, respectively.

where ξ is the distance from the lens center and $M(\xi)$ is the mass enclosed within radius ξ,

$$M(\xi) = 2\pi \int_0^\xi \Sigma(\xi')\xi'\, d\xi' \ . \tag{9.49}$$

The basic lensing geometry is illustrated in Figure 9.20. A light ray from a source S is deflected by the angle $\vec{\hat{\alpha}}$ at the lens and reaches an observer O. The angle between some arbitrarily-chosen axis and the true source position is $\vec{\beta}$, and the angle between the same axis and the image I is $\vec{\theta}$. The angular diameter distances between observer and lens, lens and source, and observer and source are denoted here as $D_{\rm d}$, $D_{\rm ds}$, and $D_{\rm s}$, respectively.

It is convenient to introduce the reduced deflection angle

$$\vec{\alpha} = \frac{D_{\rm ds}}{D_{\rm s}}\, \vec{\hat{\alpha}} \ . \tag{9.50}$$

The triangular geometry in Figure 9.20 implies that $\theta D_{\rm s} = \beta D_{\rm s} - \hat{\alpha} D_{\rm ds}$, so that the positions of the source and the image are related through the simple *lens equation*,

$$\vec{\beta} = \vec{\theta} - \vec{\alpha}(\vec{\theta}) \ . \tag{9.51}$$

The nonlinear lens equation allows for multiple images $\vec{\theta}$ at a fixed source position $\vec{\beta}$. In a flat Universe, the comoving angular-size distances simply add up, with $D_{\rm ds}(1+z_s) = D_{\rm s}(1+z_s) - D_{\rm d}(1+z_d)$.

Because of the equivalence principle, the gravitational deflection is independent of photon wavelength. In addition, since the phase space density of photons must be conserved (Liouville's theorem), gravitational lensing preserves the surface brightness of the source and only changes its apparent surface area. The total flux received from a gravitationally lensed image of a source is therefore changed in proportion to the ratio between the solid angles of the image and the source. For a circularly symmetric lens, the magnification factor μ is given by

$$\mu = \frac{\theta}{\beta} \frac{d\theta}{d\beta} \ . \tag{9.52}$$

An extended source is lensed as a sum over infinitesimal (pointlike) segments, each centered on different sky coordinates and having its own magnification factor.

9.7.1 Special Examples of Lenses

9.7.1.1 Constant Surface Density

For a mass sheet with a constant surface density Σ, equation (9.48) implies a reduced deflection angle of,

$$\alpha(\theta) = \frac{D_{\rm ds}}{D_{\rm s}} \frac{4G}{c^2 \xi} (\Sigma \pi \xi^2) = \frac{4\pi G \Sigma}{c^2} \frac{D_{\rm d} D_{\rm ds}}{D_{\rm s}} \theta \ , \tag{9.53}$$

where $\xi = D_{\rm d}\theta$. In this special case, the lens equation is linear with, $\beta \propto \theta$. Let us define a critical surface-mass density

$$\Sigma_{\rm cr} = \frac{c^2}{4\pi G} \frac{D_{\rm s}}{D_{\rm d} D_{\rm ds}} = 0.35 \, {\rm g \, cm^{-2}} \left(\frac{D}{1 \, {\rm Gpc}} \right)^{-1} \ , \tag{9.54}$$

where the effective distance D is defined through the following combination of distances

$$D = \frac{D_{\rm d} D_{\rm ds}}{D_{\rm s}} \ . \tag{9.55}$$

For a lens with $\Sigma = \Sigma_{\rm cr}$, the deflection angle is $\alpha(\theta) = \theta$, and so $\beta = 0$ for all θ. Such a lens focuses perfectly, with a single focal length. For a typical gravitational lens, however, light rays which pass the lens at different impact parameters cross at different distances behind the lens. Usually, lenses with $\Sigma > \Sigma_{\rm cr}$ somewhere in them, produce multiple images of the source.

9.7.1.2 Circularly Symmetric Lenses

For a circularly symmetric lens with an arbitrary mass profile, equations (9.48) and (9.50) give,

$$\beta(\theta) = \theta - \frac{D_{\rm ds}}{D_{\rm d} D_{\rm s}} \frac{4GM(\theta)}{c^2 \, \theta} \ . \tag{9.56}$$

A source which lies exactly behind the center of symmetry of the lens ($\beta = 0$) is imaged as a ring. Substituting $\beta = 0$ in equation (9.56) yields the angular radius of the ring to be,

$$\theta_{\rm E} = \left[\frac{4GM(\theta_{\rm E})}{c^2}\,\frac{D_{\rm ds}}{D_{\rm d}D_{\rm s}}\right]^{1/2} . \tag{9.57}$$

This so-called *Einstein radius* defines the characteristic angular scale of lensed images: when multiple images are produced, the typical angular separation between the images images is $\sim 2\theta_{\rm E}$. Also, sources which are closer than $\sim \theta_{\rm E}$ in projection relative to the lens center, experience strong lensing in the sense that they are significantly magnified, whereas sources which are located well outside the Einstein ring are magnified very little. In many lens models, the Einstein ring also represents roughly the boundary between source positions that are multiply-imaged and those that are only singly-imaged. Interestingly, by comparing equations (9.54) and (9.57) we see that the mean surface mass density inside the Einstein radius is just the critical density $\Sigma_{\rm cr}$.

For lensing by a galaxy mass M at a cosmological distance D, the typical Einstein radius is

$$\theta_{\rm E} = (0.''4)\,\left(\frac{M}{10^{11}\,M_\odot}\right)^{1/2}\,\left(\frac{D}{5{\rm Gpc}}\right)^{-1/2} . \tag{9.58}$$

9.7.1.3 Point Mass

For a point mass M the lens equation has the form,

$$\beta = \theta - \frac{\theta_{\rm E}^2}{\theta} . \tag{9.59}$$

This equation has two solutions,

$$\theta_\pm = \frac{1}{2}\,\left(\beta \pm \sqrt{\beta^2 + 4\theta_{\rm E}^2}\right) . \tag{9.60}$$

Any source is imaged twice by a point mass lens. The two images are on opposite sides of the source, with one image inside the Einstein ring and the other outside. As the source moves away from the lens (i.e. as β increases), one of the images approaches the lens and becomes very faint, while the other image approaches the true position of the source and asymptotes to its unlensed flux.

By substituting β from the lens equation (9.59) into equation (9.52), we obtain the magnifications of the two images,

$$\mu_\pm = \left[1 - \left(\frac{\theta_{\rm E}}{\theta_\pm}\right)^4\right]^{-1} = \frac{u^2+2}{2u\sqrt{u^2+4}} \pm \frac{1}{2} , \tag{9.61}$$

where u is the angular separation of the source from the point mass in units of the Einstein angle, $u = \beta\theta_{\rm E}^{-1}$. Since $\theta_- < \theta_{\rm E}$, $\mu_- < 0$, and so the magnification of the image which is inside the Einstein ring is negative implying that this image has

its parity flipped with respect to the source. The net magnification of flux in the two images is obtained by adding the absolute magnifications,

$$\mu = |\mu_+| + |\mu_-| = \frac{u^2+2}{u\sqrt{u^2+4}} \; . \tag{9.62}$$

When the source lies on the Einstein radius, we have $\beta = \theta_{\rm E}$, $u = 1$, and the total magnification becomes

$$\mu = 1.17 + 0.17 = 1.34 \; . \tag{9.63}$$

9.7.1.4 Singular Isothermal Sphere

A simple model for the mass distribution in a galaxy assumes that its collisionless particles (stars and dark matter) possess the same isotropic velocity dispersion everywhere. Surprisingly, this simple model appears to describe extremely well the dynamics of stars and gas in the cores of disk galaxies (whose rotation curve is roughly flat), as well the strong lensing properties of spheroidal galaxies.

We assume a spherically symmetric gravitational potential which confines the collisionless particles that produce it. We can associate an effective "pressure" with the momentum flux of these particles at a mass density ρ,

$$p = \rho\sigma_v^2, \tag{9.64}$$

where σ_v is the one-dimensional velocity dispersion of the particles, assumed to be constant across the galaxy. The equation of hydrostatic equilibrium (which is derived from the second moment of the collisionless Boltzmann equation) gives

$$\frac{1}{\rho}\frac{dp}{dr} = -\frac{GM(r)}{r^2} \; , \quad \frac{dM(r)}{dr} = 4\pi\, r^2\, \rho \; , \tag{9.65}$$

where $M(r)$ is the mass interior to radius r. A particularly simple solution of equations (9.64) through (9.65) is

$$\rho(r) = \frac{\sigma_v^2}{2\pi G}\,\frac{1}{r^2} \; . \tag{9.66}$$

This mass distribution is called the *singular isothermal sphere* (and will be abbreviated as SIS below). Since $\rho \propto r^{-2}$, the mass $M(r)$ increases $\propto r$, and therefore the rotational velocity of test particles in circular orbits in the gravitational potential is

$$V_{\rm c}^2(r) = \frac{G\,M(r)}{r} = 2\,\sigma_v^2 = \text{constant} \; . \tag{9.67}$$

As mentioned, the flat rotation curves of disk galaxies are naturally reproduced by this model.

By projecting the mass distribution along the line-of-sight, we obtain the surface mass density,

$$\Sigma(\xi) = \frac{\sigma_v^2}{2G}\,\frac{1}{\xi} \; , \tag{9.68}$$

where ξ is the distance from the center of the two-dimensional profile. The reduced deflection angle from (9.48),

$$\hat{\alpha} = 4\pi\,\frac{\sigma_v^2}{c^2} = (1.''16)\,\left(\frac{\sigma_v}{200\,{\rm km\,s^{-1}}}\right)^2 \; , \tag{9.69}$$

is independent of ξ and points toward the center of the lens. The Einstein radius of the SIS follows from equation (9.57),

$$\theta_{\rm E} = 4\pi \frac{\sigma_v^2}{c^2} \frac{D_{\rm ds}}{D_{\rm s}} = \hat{\alpha} \frac{D_{\rm ds}}{D_{\rm s}} = \alpha \ . \tag{9.70}$$

Due to circular symmetry, the lens equation is one dimensional. Multiple images are obtained only if the source lies inside the Einstein ring. If $\beta < \theta_{\rm E}$, the lens equation has the two solutions

$$\theta_{\pm} = \beta \pm \theta_{\rm E} \ . \tag{9.71}$$

The images at $\theta_{\pm}$, the source, and the lens all lie on a straight line. Technically, a third image with zero flux is located at $\theta = 0$; this image acquires a finite flux if the divergent density at the center of the lens is replaced by a core region with a finite density.

The magnifications of the two images follow from equation (9.52),

$$\mu_{\pm} = \frac{\theta_{\pm}}{\beta} = 1 \pm \frac{\theta_{\rm E}}{\beta} = \left(1 \mp \frac{\theta_{\rm E}}{\theta_{\pm}}\right)^{-1} \ . \tag{9.72}$$

If the source lies outside the Einstein ring, i.e. if $\beta > \theta_{\rm E}$, there is only one image at $\theta = \theta_+ = \beta + \theta_{\rm E}$. Searches for highly magnified images of faint galaxies at high redshifts are being conducted near the Einstein radius of clusters of galaxies, where the magnification factor peaks.

9.7.2 Lensing Probability

A SIS lens has the simple property that the deflection angle $\hat{\alpha}$ is independent of the impact parameter of the light ray. The condition for multiple imaging (and hence strong lensing) is then that the source would lie inside the Einstein radius. The probability that a line-of-sight to a source at a redshift z_s passes within the cross-sectional area associated with the Einstein radius of SIS lenses $\pi\theta_E^2$ gives a lensing optical depth,

$$\tau(z_{\rm s}) = \frac{16\pi^3}{H_0} \int_0^{z_s} dz \frac{D^2(1+z)^2}{(\Omega_m(1+z)^3+\Omega_\Lambda)^{1/2}} \int_0^\infty d\sigma_v \frac{dn}{d\sigma_v} \sigma_v^4, \tag{9.73}$$

where $(dn/d\sigma_v)d\sigma_v$ is the (redshift-dependent) comoving density of SIS halos with a one-dimensional velocity dispersion between σ_v and $\sigma_v + d\sigma_v$.

In calculating the probability of lensing it is important to allow for various selection effects. Lenses magnify the observed flux, and lift sources which are intrinsically too faint to be observed over the detection threshold. At the same time, lensing increases the solid angle within which sources are observed so that their number density in the sky is reduced. If there is a large reservoir of faint sources, the increase in source number due to the apparent brightening outweighs their spatial dilution, and the observed number of sources is increased due to lensing. This so-called magnification bias can substantially increase the probability of lensing for bright sources whose number-count function is steep. The magnification bias for sources at redshift z_s with luminosities between L and $L + dL$ is,

$$B(L) = \frac{1}{dn_s(L)/dL} \int_{\mu_{\rm min}}^{\mu_{\rm max}} \frac{d\mu}{\mu} \frac{dP}{d\mu} \frac{dn_s(L)}{dL}, \tag{9.74}$$

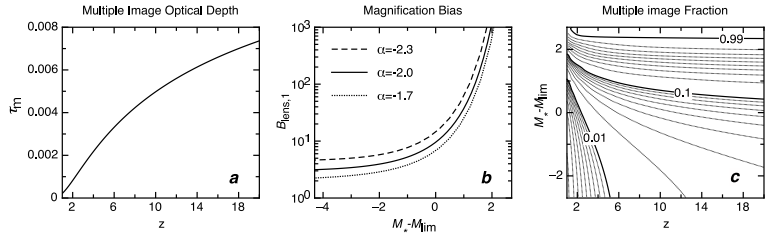

Figure 9.21 Probability for multiple imaging of high redshift galaxies by an unevolving population of SIS lenses. *Panel a*: lensing probability for obtaining multiple images τ as a function of source redshift. *Panel b*: magnification bias as a function of the difference between the characteristic magnitude of a galaxy $M_\star$ (assuming a Schechter luminosity function) and the limiting survey magnitude $M_{\rm lim}$. Three values of the faint-end slope of the luminosity function (labeled by α here) are shown. *Panel c*: Contours of the fraction of multiply-imaged sources as a function of source redshift and $(M_\star - M_{\rm lim})$, assuming a faint end slope for their luminosity function of -2. Figure credit: Wyithe, J. S. B., et al. Nature **469**, 7329 (2011)

where $n_s(< L)$ is the density of sources with luminosity $< L$ and $dP/d\mu$ is the probability for magnification μ. For example, the brighter SIS image has a magnification distribution $(dP/d\mu) = 2(\mu - 1)^{-3}$ for $2 < \mu < \infty$.

A simplified model for the redshift evolution of SIS lenses is to use the mass function of dark matter halos that was derived in §3 and identify $\sigma_v = V_{\rm c}/\sqrt{2}$ at the virial radius. Another simplified approach is to adopt the observed $(dn/d\sigma_v)$ at $z = 0$ and assume no evolution in the comoving density of lenses. The latter approach gives the approximate results shown in Figure 9.21.

Chapter Ten

The Lyman-α Line as a Probe of the Early Universe

10.1 LYMAN-α EMISSION FROM GALAXIES

We saw in §9.2.1 that young star-forming galaxies can produce very bright Lyman-α emission; indeed searching for bright line Lyman-α line emitters is one of the most effective ways to find high-z galaxies. Although the radiative transfer of these photons through their host galaxies is typically very complex, a good starting point is a simple model in which a fraction of stellar ionizing photons are absorbed within their source galaxy; the resulting protons and electrons then recombine, producing Lyman-α photons. Assuming ionization equilibrium, the rate of these recombinations must equal the rate at which ionizing photons are produced. However, direct recombinations to the ground state (which occur $\approx 1/3$ of the time, from the ratio of the case-A and case-B recombination coefficients α_A and α_B) simply regenerate the initial ionizing photon, so they do not contribute to the net balance.

Because only hot, massive stars – which live for only several million years – produce ionizing photons, it is a good approximation to assume that the rate at which any given galaxy generates these photons is proportional to its instantaneous star formation rate $\dot{M}_\star$. The proportionality constant, which we will call $N_{\rm Ly\alpha}$, depends on the initial mass function (IMF) of stars, because that determines what fraction of stellar mass enters these massive, hot stars. For example, a galaxy with a constant star formation rate, a Salpeter IMF, a metallicity $Z = 0.05\ Z_\odot$, and no binary stars produces $\dot{\mathcal{N}}_\gamma = 4.3\times10^{53}$ ionizing photons per second per $M_\odot\ {\rm yr}^{-1}$ in stars formed. However, a top-heavy Population III IMF has an ionization efficiency that is larger by more than an order of magnitude.

Finally, if we assume as usual that $f_{\rm esc}$ of the ionizing photons escape their host galaxy, then the intrinsic line luminosity of a galaxy is

$$L_{\rm Ly\alpha}^{\rm int} = \frac{2}{3}\dot{\mathcal{N}}_\gamma h\nu_\alpha(1-f_{\rm esc})\dot{M}_\star. \tag{10.1}$$

For context, a Salpeter IMF from 1 to 100 $M_\odot$ with $Z = 0.05Z_\odot$ has a prefactor $4.4\times10^{42}(1-f_{\rm esc})$erg s^{-1}, if the star formation rate $\dot{M}_\star$ is measured in $M_\odot$ yr^{-1}.

Unfortunately, inferring physical properties about distant galaxies from the Lyman-α line is complicated not only by the uncertain factors $f_{\rm esc}$ and N_γ but also by the radiative transfer of these line photons through the interstellar and circum-galactic medium surrounding each galaxy. Because the Lyman-α line is so optically thick in both the galaxy's ISM and the nearby IGM, these photons scatter many times before they can escape the galaxy, and once they leave it they can be scattered away

from the line of sight and vanish. This scattering can change not only the overall brightness of the line but also its frequency structure and relation to the galaxy's continuum photons. The *observed* line luminosity is then

$$\boxed{L_{\rm Ly\alpha}^{\rm int} = \frac{2}{3} T_{\rm Ly\alpha}^{\rm IGM} T_{\rm Ly\alpha}^{\rm ISM} \dot{\mathcal{N}}_\gamma h\nu_\alpha (1 - f_{\rm esc}) \dot{M}_\star,} \tag{10.2}$$

where $T_{\rm Ly\alpha}^{\rm ISM}$ is the fraction of Lyman-α photons that are transmitted through the galaxy's ISM and $T_{\rm Ly\alpha}^{\rm IGM}$ is the fraction transmitted through the IGM.

10.1.1 Radiative Transfer of Lyman-α Photons Through the Interstellar Medium

We will first consider radiative transfer within the galaxy and its immediate environs; we will defer discussion of IGM scattering until §10.2. The important difference from continuum transfer is that line photons can scatter many times (changing both their direction and frequency) as they traverse the ISM. In the case of Lyman-α photons, scattering cannot destroy them (unless collisions mix the $2s$ and $2p$ states, which requires much higher densities than typical for the IGM), but dust absorption can. Depending on the geometry of the ISM, the increased path length can increase or decrease the brightness of the line relative to the continuum.

Some simple toy models of radiative transfer help to develop some intuition for this situation. We will generally assume that the absorption cross-section follows the usual *Voigt profile*, $\sigma_\alpha = \sigma_0 \phi_V$, which allows for both thermal broadening (which causes gaussian broadening in the core of the line) and natural Lorentzian broadening in the wings (which comes from the finite lifetime of the upper state). This may be written as a convolution of these two mechanisms,

$$\phi_V(\nu) = \int_{-\infty}^{\infty} M(v) L[\nu(1 - v/c)] dv, \tag{10.3}$$

where the integral is over the line-of-sight thermal velocities of the particles. Here $M(v)$ is the Maxwell-Boltzmann distribution at a temperature T,

$$M(v)dv = \frac{1}{\sqrt{\pi b^2}} e^{-v^2/b^2} dv, \tag{10.4}$$

with $b = \sqrt{2k_B T/m_p}$ parameterizing the thermal broadening.[i] L is the natural line profile, which for Lyman-α is given by equation (4.8). This is often approximated by a Lorentzian function,[ii]

$$L(\nu) \approx \frac{1}{\pi} \frac{\gamma}{(\nu - \nu_\alpha)^2 + \gamma^2}, \tag{10.5}$$

where $\gamma = A_{21}/(4\pi)$ is the decay constant. In this approximation, the Voigt profile can be written

$$\phi_V(x)dx = \frac{1}{\sqrt{\pi}} \left(\frac{\nu_\alpha}{\nu}\right) V(x) dx, \tag{10.6}$$

[i] The Doppler parameter can include turbulence as well by adding the turbulent velocity in quadrature.

[ii] However, in cosmology the optical depth can be so high that the asymmetry of the full profile is visible.

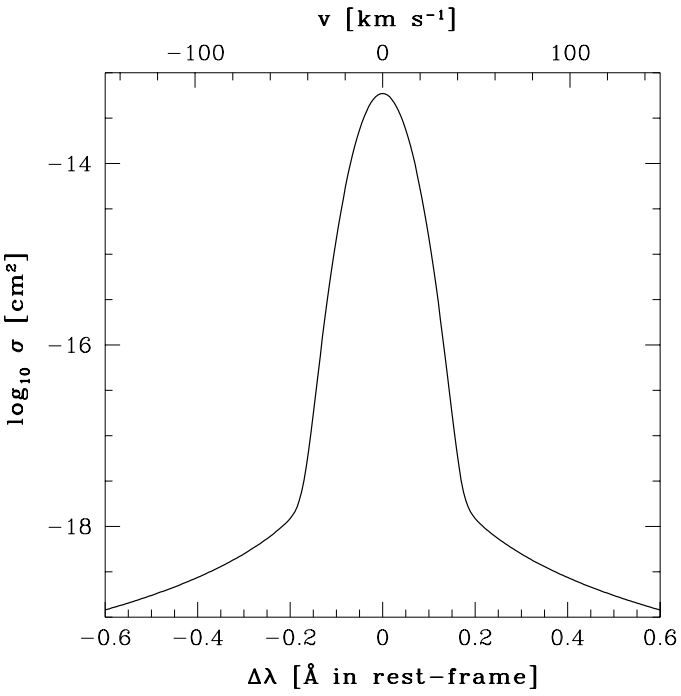

Figure 10.1 Cross-section for Lyman-α absorption, as a function of wavelength offset from line center (bottom axis) or velocity difference (top axis). We include thermal and natural broadening generated by gas with $T = 10^4$ K. Figure credit: Santos, M.R. 2004, MNRAS, 349, 1137.

where $x = (\nu - \nu_\alpha)/\nu_D$ is the normalized frequency, with a Doppler broadening $\nu_{\rm D}/\nu_\alpha = b/c$ and

$$V(x) = \frac{A(x)}{\pi} \int_{-\infty}^{\infty} dy \frac{e^{-y^2}}{[B(x) - y]^2 + A^2(x)}, \tag{10.7}$$

and finally $A(x) = (\gamma/\nu_D) \times (\nu_\alpha/\nu)$ and $B(x) = x(\nu/\nu_\alpha)$. An approximation to the Voigt function makes the line structure apparent:

$$V(x) \approx e^{-B^2} + \frac{1}{\sqrt{\pi}} \frac{A}{A^2 + B^2}. \tag{10.8}$$

We will be particularly interested in the profile far from the line core (where $B \gg 1$). There, $\sigma_\alpha \approx (\gamma/\nu_D)/(\sqrt{\pi}x^2)$ where $\gamma/\nu_D 4.72 \times 10^{-4} T_4^{-1/2}$ and $T_4 = (T/10^4\ {\rm K})$. Figure 10.1 shows the absorption cross-section for absorbing gas with $T = 10^4$ K (including the full line broadening); note the Gaussian core with width ~ 10km s^{-1} and the much weaker, but broader, damping wings.

- *Homogeneous, static H I slab, moderately optically thick:* First consider a Lyman-α photon produced inside a homogeneous static medium of pure H I, with total line-center optical depth $\tau_0 \gg 1$; note that because τ_0 is proportional to distance in the medium, we can use it as a proxy for physical location

within the system. So long as the photon remains in the Doppler core of the line, it barely diffuses spatially before being scattered by an atom. When a line photon of frequency $x_{\rm in}$ is absorbed by an atom, it re-emits a line photon of the same frequency in its own rest frame. However, in an observer's frame there will be a net frequency shift determined by the Lorentz transformation between the frames. To linear order, this is

$$x_{\rm out} \approx x_{\rm in} - \frac{\mathbf{v}_a \cdot \mathbf{k}_{\rm in}}{v_{\rm th}} + \frac{\mathbf{v}_a \cdot \mathbf{k}_{\rm out}}{v_{\rm th}} + g(\mathbf{k}_{\rm in} \cdot \mathbf{k}_{\rm out} - 1), \tag{10.9}$$

where $\mathbf{v}_a$ is the velocity vector of the atom, $v_{\rm th} = (2k_BT/m_p)^{1/2}$ is the thermal velocity of the gas, and $\mathbf{k}_{\rm in}$ and $\mathbf{k}_{\rm out}$ are the propagation directions of the incoming and outgoing photons, respectively. The last term g represents recoil; it is unimportant here, but we will revisit it in §11.2.2. Typically, the scattering atom will have the same velocity along the photon's direction of motion as the atom that emitted it, but it can have a much larger total velocity. In that case, the scattered photon will be far from the line center.

If the medium is not too optically thick, so that the damping wings are themselves optically thin, the resulting photon can escape so long as it is produced at a frequency where $\tau(x) < 1$; for $\tau_0 = 10^3$, this corresponds to $x \sim 2.6$. We therefore generically expect that the resulting emission will have a double-peaked profile: photons near line-center do not escape; only when they diffuse to large positive or negative velocity are they able to escape. The Lyman-α surface brightness distribution will also be compact, because photons escape after a single scattering.

- *Homogeneous, static H I slab, very optically thick:* In a moderately optically thick medium, these escaping photons simply result from rare scatterings off high-velocity atoms. If, on the other hand, the damping wings are optically thick, $\tau_{0a} > 10^3$, so that once a photon is scattered into the wing the next scattering is more likely to be from interaction with an atom in the wings of its line than with an atom traveling at a matched velocity, the problem is more complicated, though the net result is easy to understand: the photons must make it even farther into the wings to escape.

  To do so, they must undergo a random walk of repeated scatterings, which occasionally take them far enough from line center to escape. Because scatterings usually occur in the core, each one induces an rms frequency shift $x \sim 1$, with a small bias $-1/x$ toward returning to line center; a photon thus typically undergoes $N_s \sim x^2$ scattering events before returning to line center. Between scatterings, the photon traverses a path length (in optical depth units) of $\tau\Phi(x) \sim 1$. Thus, over its entire random walk, it diffuses a distance of $\tau_0^{\rm rms} \sim \sqrt{N_s}\tau \sim |x|/\phi_V$. If this distance exceeds the size of the system (τ_0 in these units), the photon can escape. In the wings of the line where $\phi_V \sim a/x^2$, this requires that the photon have a critical normalized escape frequency

$$|x_{\rm esc}| \sim (a\tau_0)^{1/3} \approx 30 T_4^{-1/3} N_{21}^{1/3}, \tag{10.10}$$

where N_{21} is the column density of the system in units of 10^{21} cm^{-2}. Thus, in this highly-optically thick case, the photons must scatter far enough in the wings of the line to physically escape the system before scattering back to line center. This, combined with the power law form of ϕ_V in the wings, also makes the blue and red emission peaks wider than than in the moderate optical depth case. The surface brightness of the line will be extended even if the source is compact, because photons diffuse spatially as well as in frequency before escaping.

- *Homogeneous H I slab, with velocity gradient:* We next consider a medium with a velocity gradient. Such a gradient can either correspond to expansion, arising from winds (which we believe to be ubiquitous in the star-forming galaxies likely to host Lyman-α emission lines), or contraction, from the infall of surrounding material around the galaxy.

  First consider an expanding medium. Then, according to equation (10.9), scattered photons typically obtain a redshift: $\mathbf{v}_a \cdot \mathbf{k}_{\rm in}$ is positive for photons propagating outward, while $\langle \mathbf{v}_a \cdot \mathbf{k}_{\rm out} \rangle = 0$, so $x_{\rm out} < x_{\rm in}$ on average. Photons with $x < 0$ are therefore moved farther into the line wings, facilitating their escape, while photons with $x > 0$ are moved back toward line center. So long as the expansion velocity is much larger than the thermal velocities, this will prevent photons that experience large positive frequency jumps from escaping. Thus, we expect only a single emission line on the red side. In contrast, in a contracting medium photons typically obtain a blueshift, producing a single emission line on the blue side.

  In this case the frequency shift of the surviving line depends upon the velocity and density structure of the medium. The case of most practical relevance is a wind, in which a large column of H I occurs at $\pm v_{\rm wind}$ along the line of sight, with negligible absorption elsewhere. In this case photons that begin their escape toward the observer (i.e., through the blueshifted wind) are absorbed. After their first scattering, photons that begin their escape toward the far component of the wind lie to the red side of the line. Those that scatter back toward the observer are then far to the red of the (blueshifted) wind and can continue to the observer. The observed velocity offset is then $v_{\rm wind}$ and provides a good diagnostic of the wind velocity.

- *Homogeneous H I slab with dust:* Now we can add dust to a (static) medium and see how it can destroy the Lyman-α photons. We let the *total* dust interaction cross-section, per hydrogen nucleus, be σ_d; this includes both absorption, with a cross-section $\sigma_a = \epsilon_a \sigma_d = (\sigma_{a,21}/10^{-21}\ {\rm cm}^2)$ per hydrogen atom, and scattering. For the well-studied dust in the Milky Way, $\sigma_{a,21} \approx 1$ and $\epsilon_a \approx 0.5$; of course this will depend on the metallicity and dust formation mechanisms in high-redshift galaxies (see §9.6.2). The average absorption probability per interaction (with either dust or H I) is therefore

$$\epsilon = \frac{\sigma_a}{x_{\rm HI}\phi_V(x)\sigma_0 + \sigma_d} \approx \frac{\beta}{\phi_V(x)}, \tag{10.11}$$

where $\beta = \sigma_a/(x_{\rm HI}\phi_V = 1.69 \times 10^{-8} T_4^{1/2} \sigma_{a,21}/x_{\rm HI}$ and we have assumed that dust interactions are rare compared to H I scattering.

Now recall that, in order to escape the H I, the photon must first scatter far into the wings of the line and then stay in the wings as it spatially diffuses out of the system. During that process, the photon will scatter N_s times; the probability that it is absorbed is therefore $P_{\rm abs} \sim N_s\epsilon \sim x^4\beta/a$ in the damping wing. This is near unity if $|x| > x_{\rm abs}$, where

$$x_{\rm abs} \sim (a/\beta)^{1/4} \sim 12.9 \left(\frac{x_{\rm HI}}{T_4 \sigma_{a,21}} \right)^{1/4}. \tag{10.12}$$

A typical photon will therefore be unable to escape if $x_{\rm esc} > x_{\rm abs}$; if the line center optical depth exceeds a critical value $\tau_c \sim (a\beta^3)^{1/4}$, the emission line will be strongly suppressed. This corresponds to a column density of only $N_{21,c} = 0.08 T_4^{1/4} (x_{\rm HI}/\sigma_{a,21})^{3/4}$, well below the typical column densities of galaxies (which are comparable to damped Lyman-α absorbers [DLAs]). Thus Lyman-α absorption can be very important inside the ISM. In general, in a uniform medium the line photons are more affected by dust than continuum photons, because the many scatterings they suffer forces them to have a much longer path length than continuum photons, providing a much larger opportunity for dust absorption.

- *Multiphase medium with dust:* Finally, we consider a medium in which both the H I and dust are confined to optically thick, discrete clouds separated by a highly-ionized, dust-free "inter-cloud medium." Here the results will clearly depend on the geometry of the system, but some general considerations do apply. First, note that an inhomogeneous medium will allow *more* transmission than a homogeneous slab with identical column density of neutral gas, because of the same arguments we saw for transmission in an inhomogeneous IGM (see §4.3.2). Moreover, the line photons can be *less* affected by dust than continuum photons, because the line photons scatter off the *surface* of the clouds, while the continuum photons plow through them and can encounter *more* dust.

  Detailed calculations show that the frequency shift necessary for dust absorption to dominate over resonant scattering in the line wings, $x_{\rm abs}$, is similar in magnitude to the homogeneous case. However, dust was so important in that example because Lyman-α photons *needed* to diffuse in frequency in order to escape the medium. This is not the case for a multiphase medium. In this case, photons enter each cloud on their surface and suffer relatively few scattering events inside each cloud before spatially diffusing back out. They can then travel a large distance before hitting another cloud, and spatial diffusion through the inter-cloud medium provides most of the impetus toward escape. Thus, dust absorption will be relatively weak provided that the typical frequency shift before escape is less than $x_{\rm abs}$.

  In this case, photons obtain frequency shifts both from the thermal motions of the scattering atoms and from the velocity dispersion between the absorbing clouds; if the latter is large (as would be the case if most of the dust

were buried in dense molecular clouds), it dominates the frequency diffusion, because – just as for a wind – each cloud is so optically thick that in the observer's frame the photon leaves each cloud with a velocity offset corresponding to that cloud's velocity. If the clouds have a large velocity dispersion, then dust absorption within each cloud will dominate over resonant scattering, because the photons will enter each cloud in the wings of the line.

Although each of these toy models is obviously much simpler than a real galaxy, together they illustrate the complexity of the radiative transfer problem and the many parameters that can dramatically affect the Lyman-α line's amplitude and shape, as well as the surface brightness of a line emitter. In general, even discounting uncertainties from IGM transmission discussed below, the Lyman-α line is therefore typically very difficult to interpret and is not regarded as, for example, a very reliable measure of the star formation rate. However, its brightness in many galaxies makes it such a useful signpost that it is still the subject of intense study.

10.1.2 Other Emission Lines

Because it can be such a bright line, and because its ultraviolet rest wavelength redshifts it into the optical or near-infrared in distant galaxies, the Lyman-α line gets the lion's share of attention. But other emission lines can be as or even more useful for certain diagnostics, and we briefly mention them here. This is of course a very extensive field of research, and so we refer the interested reader to the literature and other textbooks for more information (see the Appendix).

1. *Other hydrogen lines:* The other Lyman-series lines are almost never visible in high-redshift galaxies; after several scatterings, these photons are "recycled" via radiative cascades into either Lyman-α photons or a pair of photons from the forbidden $2s \rightarrow 1s$ decay (see §11.2.2). However, Balmer-series photons (and those beginning at even higher levels) are very useful diagnostics. They are initially generated through the same process as Lyman-α – recombinations following ionizations near hot, massive stars – but because such photons can only interact with atoms already in the $n = 2$ state, they are not subject to scattering in the interstellar medium and escape galaxies relatively easily (especially since they have relatively long wavelengths and so are less subject to dust absorption, e.g., the Hα line lies at 6563 Å). They therefore offer much more robust measures of star formation rates, subject only to the uncertainty in the IMF.

   Unfortunately, although Hα is extremely important for low-redshift galaxies, its relatively long rest wavelength has so far limited its usefulness for high-redshift galaxies.

2. *Helium lines:* He II has the same electronic structure as H I, but shifted to four times larger energies. As a result, its ionization potential is well beyond the cutoff of most stars – only rare Wolf-Rayet stars (i.e., massive stars undergoing rapid mass loss) and the most massive Population III stars are hot enough to significantly ionize it. He II Balmer-α photons (with a rest

wavelength of 1640 Å) are therefore the most promising diagnostic of such massive stars: they are produced through recombination cascades following the ionization of He II.

3. *Metal lines:* In nearby galaxies, many metal lines offer diagnostics of ISM characteristics like the density, metallicity, and temperature of the nebulae surrounding star-forming regions. As instruments improve, these will no doubt be just as useful for measurements of high-z galaxies, although (with most of the lines having rest wavelengths in the optical) they are less accessible for the more distant sources.

10.2 THE GUNN-PETERSON TROUGH

We now briefly discuss the fate of photons that begin blueward of Lyman-α during the reionization era. These photons will redshift through the IGM; if they should pass through the Lyman-α resonance, they will experience substantial absorption from that gas. The scattering cross-section of the H I Lyman-α resonance line is given by equation (4.8), and we have already computed the total optical depth for a photon that redshifts through the Lyman-α resonance as it travels through the IGM, the so-called Gunn-Peterson optical depth in equation (4.13). The most important aspect of this calculation is the enormous overall optical depth in a fully-neutral IGM, $\tau_\alpha \sim 6.5 \times 10^5 x_{\rm HI}$ at $z \sim 9$. Thus we expect that, before reionization, photons that redshift across the Lyman-α transition will be completely extinguished (and, indeed, the same will be true so far as the ionized fraction is $< 10^{-3}$).

However, not all photons will redshift through the resonance during the reionization era. Suppose that a photon is emitted by a source at a redshift z_s beyond the "redshift of reionization" $z_{\rm reion}$, which for the purposes of this calculation is simply the last redshift along the particular line of sight of interest where $x_{\rm HI} = 1$. (Note that this differs from the conventional definition of the end of reionization as the moment of "overlap" between the ionized bubbles; the variations along different lines of sight can themselves contain interesting astrophysical information.) For simplicity, we will further assume that $x_{\rm HI} = 1$ for all $z > z_{\rm reion}$. The corresponding scattering optical depth of a uniform, neutral IGM is a function of the observed wavelength $\lambda_{\rm obs}$,

$$\tau_\alpha(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz \frac{cdt}{dz} n_{\rm H,0}(1+z)^3 \sigma_\alpha \left[\nu_{\rm obs}(1+z)\right]. \qquad (10.13)$$

At wavelengths corresponding to the Lyman-α resonance between the source redshift and the reionization redshift, $(1+z_{\rm reion})\lambda_\alpha \leq \lambda_{\rm obs} \leq (1+z_s)\lambda_\alpha$, the optical depth is given by equation (4.13). Since $\tau_\alpha \sim 10^5$, the flux from the source is entirely suppressed in this regime. Similarly, the Lyman-β resonance produces another trough at wavelengths $(1+z_{\rm reion})\lambda_\beta \leq \lambda_{\rm obs} \leq (1+z_s)\lambda_\beta$, where $\lambda_\beta = (27/32)\lambda_\alpha = 1026\,$Å, and the same applies to the higher Lyman series lines. If $(1+z_s) \geq 1.18(1+z_{\rm reion})$ then the Lyman-α and the Lyman-β resonances overlap and no flux is transmitted between the two troughs. The same holds for the higher Lyman-series resonances down to the Lyman limit wavelength of $\lambda_c = 912$Å.

At wavelengths shorter than λ_c, the photons may be absorbed when they photoionize atoms of hydrogen or helium, even if they do not redshift into the Lyman series lines. The bound-free absorption cross-section of hydrogen is given by equation (4.16); the appropriate parameters for He II are given in §4.4.4 as well. A reasonable approximation to the total cross-section for a mixture of hydrogen and helium with cosmic abundances in the range of $4h\nu_{\rm H,0} = 54.4\,{\rm eV} < h\nu < 10^3$ eV is $\sigma_{bf} \approx \sigma_0(\nu/\nu_{\rm H,0})^{-3}$, where $\sigma_0 \approx 6 \times 10^{-17}\,{\rm cm}^2$. The redshift factor in the cross-section then cancels exactly the redshift evolution of the gas density and the resulting optical depth depends only on the elapsed cosmic time, $t(z_{\rm reion}) - t(z_s)$. At high redshifts this yields,

$$\tau_{bf}(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz \frac{cdt}{dz} n_0 (1+z)^3 \sigma_{\rm bf}\left[\nu_{\rm obs}(1+z)\right]$$
$$\approx 1.5 \times 10^2 \left(\frac{\lambda_{\rm obs}}{100{\rm \AA}}\right)^3 \left[\frac{1}{(1+z_{\rm reion})^{3/2}} - \frac{1}{(1+z_s)^{3/2}}\right] \quad (10.14)$$

The bound-free optical depth only becomes of order unity in the extreme UV to soft X-rays, around $h\nu \sim 0.1$ keV, a regime which is unfortunately difficult to observe due to absorption by the Milky Way galaxy.

Together, these effects imply very strong absorption of nearly all photons that begin blueward of $\lambda_\alpha(1+z_r)$, except for a recovery at very short wavelengths and the gaps between the Lyman-series troughs (though these will be blanketed by the Lyman-α and other forests just below $z_{\rm reion}$, so even they will be extremely optically thick).

10.3 IGM SCATTERING IN THE BLUE WING OF THE LYMAN-α LINE

We now return to the fate of photons emitted within (or near) the Lyman-α line of a galaxy or quasar. In this case, the relative velocity and broadening of the line from bulk, thermal, or turbulent motions is very significant, because it determines whether the photons pass through the Lyman-α resonance – and so experience the full Gunn-Peterson absorption – or remain redward of line center, experiencing much less absorption. We also must consider the environment of the source: whether it is embedded in completely neutral gas or in an ionized bubble, and the surrounding velocity field. In this section we will focus on photons emitted blueward of, but still near to, line center.

10.3.1 Resonant Scattering Inside Ionized Bubbles

Photons that begin slightly blueward of line center redshift into the Lyman-α resonance near to their source. In most models, this nearby region will already have been ionized, either by the source itself or by its neighbors (if it is part of a much larger ionized bubble). Thus it may seem that these photons will survive their journey through the IGM.

However, if we recall that $\tau_\alpha > 10^5 x_{\rm HI}$ at these redshifts, it is immediately apparent that even in highly-ionized media the absorption can be substantial. In

practice, the short mean free paths at high redshifts will most likely prevent the gas from becoming extremely ionized. We can estimate the residual ionized fraction inside an H II region in which the comoving mean free path is λ (which could be restricted either by the ionized bubble walls or by LLSs) by assuming ionization equilibrium and a uniform emissivity (or in other words that each bubble contains many sources). The equilibrium condition is then $\Gamma n_{\rm HI} = \alpha_B n_e n_{\rm H}$, with $\Gamma \approx \epsilon_{\rm ion}\bar{\sigma}_{\rm HI}\lambda/(1+z)$, $\epsilon_{\rm ion}$ the proper emissivity of ionizing photons (by number), and $\bar{\sigma}_{\rm HI} \sim 2\times10^{-18}$ cm^2 the frequency-averaged cross-section. If we use the simplest model for the ionizing sources, in which the rate of ionizing photon production is proportional to the rate at which gas accretes onto galaxies, we can write (see equation 8.2) $\epsilon_{\rm ion} = \zeta \dot{f}_{\rm coll} n_H$. But we also know $Q_{\rm HII} = \zeta f_{\rm coll}/(1+\bar{n}_{\rm rec})$, where $\bar{n}_{\rm rec}$ is the mean number of recombinations per atom. So we can rewrite the ionizing efficiency ζ in terms of the overall ionized fraction and solve for the resonant optical depth due to residual neutral gas $x_{\rm HI}$ inside the bubble:

$$\tau_\alpha^{\rm res}(\delta) \approx 40 \frac{(1+\delta)^2}{Q_{\rm HII}(1+\bar{n}_{\rm rec})} \left(\frac{10\,{\rm Mpc}}{\lambda}\right) \left(\frac{f_{\rm coll}}{df_{\rm coll}/dz}\right), \tag{10.15}$$

where we have assumed that the IGM is isothermal for the recombination coefficient. The factor involving the collapsed fraction is typically of order a few.

Clearly the optical depth for these photons is large in realistic models; note, however, that it is small enough that many of the radiative transfer effects important for photon escape from galaxies are not important, and the absorption from each gas parcel will not have a large frequency width.

10.3.2 The Proximity Effect and Quasar "Near-Zones"

We have now found that an average location inside an ionized bubble is not likely to be ionized strongly enough to allow significant transmission before reionization. However, the region immediately outside the of an ionizing source will be more ionized than average thanks to photons from that source. At moderate redshifts, this "proximity effect" is a useful measure of the ionizing background, and it is a very attractive probe of the reionization era as well.

The profile of the ionization rate around a quasar at moderate redshifts is simple to understand. Suppose that there is a uniform metagalactic background with amplitude $\Gamma_{\rm bg}$. The central quasar with a luminosity L_ν produces a specific intensity $J_\nu \propto L_\nu/R^2$, where R is the distance from the quasar. Thus, we expect an ionization rate $\Gamma_q = \Gamma_{q,0}/R^2$. Assuming ionization equilibrium, we then have

$$\tau(R) \propto (\Gamma_{\rm bg} + \Gamma_{q,0}/R^2)^{-1}. \tag{10.16}$$

A simple fit to the absorption profile as a function of distance from the quasar is sufficient for deriving $\Gamma_{\rm bg}$, especially if $\Gamma_{q,0}$ can be estimated from the observed luminosity of the quasar redward of the Lyman-α line. In practice, these estimates are complicated by variations in the Lyman-α forest lines themselves and by the biased environments of quasars: the quasar will only induce substantial changes in the radiation field within a compact "proximity zone" around the quasar where

$\Gamma_q > \Gamma_{\rm bg}$. This corresponds to

$$R_{\rm prox} = \frac{1.2\Gamma_{12}^{-1/2}}{\alpha+3}\left(\frac{\nu L_\nu}{10^{44}\ {\rm erg\ s^{-1}}}\right)^{1/2}\ {\rm proper\ Mpc}, \qquad (10.17)$$

where α is the quasar spectral index and L_ν is evaluated at the H I ionization edge. This places the proximity zone within the overdense environment of the quasar's halo; the increased absorption from this excess gas partially cancels the effect of the increased ionizing background, making the proximity effect more difficult to see.

Because the ionizing background is much smaller during the reionization era, it may at first appear that the proximity effect will be easier to observe. However, in reality the effect is much more difficult to interpret because the IGM is so optically thick. In this situation, the observable pattern near a luminous source will be gradually increasing absorption until saturation is reached. Figure 10.2 shows some examples; the curves here have each been averaged over several independent lines of sight to reduce the scatter from the inhomogeneous IGM. Here the horizontal dotted line marks 10% transmission; this is conventionally used to mark the edge of the transmission region.

The key point is that, during the reionization era, there are two possible reasons why such saturation can occur. The first is if the source (usually a quasar) is still in the process of ionizing its neutral surroundings. Then there will be a sharp transition between the highly-ionized H II region and the nearly neutral gas at its edge, which will manifest itself as a dramatic increase in the local optical depth. The second is more similar to the classical proximity effect, except that the absorption may saturate long before the local ionization rate reaches the background value. Because the observed edge of the transmission does not necessarily correspond to the classical proximity zone, this feature is usually referred to as the "near-zone."

In the first case, the size of the H II region depends on the ionizing luminosity of the quasar (which can be estimated from the spectrum redward of Lyman-α), the age of the quasar t_Q, and the average neutral fraction before the quasar appeared, $\bar{x}_{\rm HI}$. The basic radiative transfer problem has already been solved in SS8.2 and 8.9.2; for the purposes of a simple estimate, if recombinations can be neglected, the proper radius of the H II region is (c.f. equation 8.3)[iii]

$$R_b \approx \frac{4.2}{\bar{x}_{\rm HI}^{1/3}}\left(\frac{\dot{N}}{2\times10^{57}\ {\rm s^{-1}}}\right)^{1/3}\left(\frac{t_Q}{10^7\ {\rm yr}}\right)^{1/3}\left(\frac{1+z}{7}\right)^{-1}\ {\rm Mpc}, \qquad (10.18)$$

where $\dot{N}_Q$ is the rate at which the quasar produces ionizing photons and we have assumed that all the ionizing photons are absorbed but ignore secondary ionizations. Note that $R_b \propto (\dot{N}_Q t_Q/\bar{x}_{\rm HI})^{1/3}$, varying relatively slowly with these parameters.

However, the absorption can become saturated well before this limit is reached. To estimate this, we suppose that the edge of the near-zone is where the transmission falls below $T_{\rm lim}$, or the optical depth rises above $\tau_{\rm lim}$. We adopt $T_{\rm lim} = 0.1$

[iii] Here we ignore relativistic effects in the expansion, which are important at early times. See equation (8.11) for a more accurate expression.

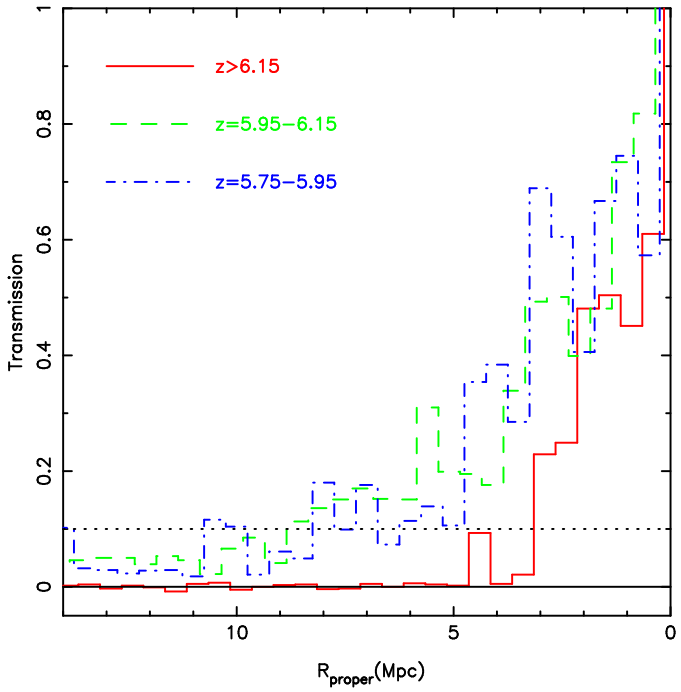

Figure 10.2 Average absorption profiles near the Lyman-α line for quasars in three different redshift bins. Note that the Lyman-α emission lines have been fitted and removed. The three redshift bins average over 8 ($5.75 < z < 5.95$), 9 ($5.95 < z < 6.15$), and 4 ($z > 6.15$) quasars. The horizontal dotted line marks 10% transmission, conventionally taken as the edge of the near-zone. Figure credit: Carilli, C.L. et al. 2010, ApJL, 714, 834.

as a fiducial value (comparable to existing observations). Assuming that the background ionization rate can be neglected (likely a good assumption at these very high redshifts), the transmission reaches this limiting value at a proper radius

$$R_{\rm lim} \approx 3.1 \left(\frac{\dot{N}}{2\times10^{57}\ {\rm s}^{-1}}\right)^{1/2} \left(\frac{T}{2\times10^{4}\ {\rm K}}\right)^{0.38} \left(\frac{\tau_{\rm lim}}{2.3}\right)^{1/2} \left[\frac{3\alpha}{(\alpha+3)}\right]^{1/2} \left(\frac{1+z}{7}\right)^{-9/4}, \quad (10.19)$$

where the T dependence enters through the recombination coefficient for ionization equilibrium. Note that this limiting radius is independent of the neutral fraction of the material outside of the ionized zone, and it is slightly more sensitive to the quasar luminosity, $R_{\rm lim} \propto \dot{N}_Q^{1/2}$.

There are two caveats on these size estimates. First, equation (10.19) can only apply if the quasar bubble has reached that size. This requires

$$t_Q > 4.2\times10^{6} \bar{x}_{\rm HI} \left(\frac{R_{\rm lim}}{3.1\ {\rm Mpc}}\right)^{3} \left(\frac{\dot{N}}{2\times10^{57}\ {\rm s}^{-1}}\right)^{-1} \left(\frac{1+z}{7}\right)^{3} {\rm yr}. \quad (10.20)$$

(Adding recombinations and clumping will increase this scale by a factor of no more than a few.) Interestingly, this timescale is comparable to the canonical quasar lifetime $t_Q \sim 10^7$ yr in fully neutral gas, but for quasars positioned near the end of reionization (which are actually accessible to observations) it is very short.

Moreover, our expressions for R_b and $R_{\rm lim}$ implicitly ignore the possibility of LLSs or even denser regions in the IGM. If the quasar radiation encounters a highly overdense region that can maintain $\tau > 1$ in ionization equilibrium, the ionizing radiation will be highly attenuated at larger distances. Although these systems are likely to be rare near the quasar (where the radiation field is particularly strong), they are difficult to identify in the highly saturated forest spectra found during reionization, and they present an important systematic concern for measurements of "near zones."

We therefore expect most quasar near-zones be limited by the proximity effect rather than the bubble size. If so, these zones can tell us little about the ionization state of the surrounding gas. In principle, this supposition can be tested by examining the luminosity dependence of the near-zone size, although the modest variation between the two models, and the large scatter intrinsic to any measurement in an inhomogeneous IGM, has made differentiating them difficult to date. Figure 10.3 shows the measured near-zone sizes for a number of quasars at $z > 5.75$. The left panel shows the trend with redshift (here all the near-zone sizes have been normalized to a common luminosity using the $R_b \propto \dot{N}_Q^{1/3}$ relation), while the right panel shows the dependence on absolute magnitude (with the mean trend with redshift removed.

In the right panel, the dotted curve shows $R_b \propto \dot{N}_Q^{1/3}$ (with arbitrary scaling); this is not a fit but is shown only for illustrative purposes. Clearly the large scatter in the near-zone sizes, even after a simple redshift correction, make it difficult to distinguish this behavior from that expected for the more classic proximity effect, $R_b \propto \dot{N}_Q^{1/2}$.

Nevertheless, there is clearly a steady increase in the near-zone size as redshift declines. One possible interpretation is a decrease in $\bar{x}_{\rm HI}$ with cosmic time; the data would require a decline by ~ 10 over the range $z = 6.4$ to $z = 5.8$. However, presuming that $z \sim 6$ is the tail end of reionization, the proximity effect is more likely to fix the near-zone size. In that case, the trend with redshift is most likely attributable to a rapid increase in the background ionization rate (by a factor of > 3), which can substantially boost the total ionization rate in the outskirts of the quasar's proximity zone.

Currently, the most challenging aspect of this measurement – other than finding these quasars in the first place – is determining the quasar's location. The only tools we have are the redshifts of the source's emission lines. Unfortunately, most quasars have strong internal motions and winds, which displaces many of the emission lines from the systemic redshift of the host. The best choices are low-excitation lines (such as Mg II) or, even better, lines from the host galaxy itself. Any such lines in the optical or UV are overwhelmed by the quasar's own emission, so the most useful lines turn out to be those of CO, which are strong in these rapidly star-forming galaxies.

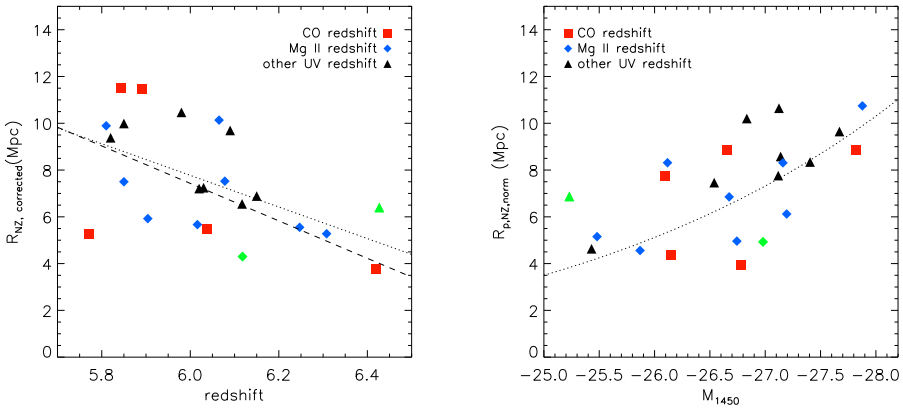

Figure 10.3 *Left:* Measured radii of near zones in a set of high-z quasars; the symbols denote the method used to compute the quasar's redshift. All near-zone measurements have been scaled to a common quasar luminosity using the R_b relation in equation (10.18) to better illustrate the trend with redshift. Typical errors in the near-zone size are ~ 1 Mpc. The two lines are fits to the trend with redshift. *Right:* Dependence of the near-zone size on quasar absolute magnitude; all the data points have been scaled to a common redshift using the mean relation in the left-hand panel to better illustrate the behavior with luminosity. The dotted line shows $R_b \propto \dot{N}_Q^{1/3}$ with arbitrary scaling; note that it is not a fit but is merely meant to guide the eye. Figure credit: Carilli, C.L. et al. 2010, ApJL, 714, 834.

There is one additional, and very attractive, way to differentiate between R_b and $R_{\rm lim}$: by examining the absorption in higher Lyman-series lines. Because $R_{\rm lim}$ depends on the maximum detectable optical depth $\tau_{\rm lim}^{1/2}$, it will increase by the square root of the optical depth ratio between different lines; for Lyman-β, this means $R_{\rm lim}^{\beta} \approx 2.5 R_{\rm lim}^{\alpha}$. However, at the edge of the ionized bubble the neutral fraction presumably increases by an enormous amount over a very small distance, so both Lyman-α and Lyman-β should become optically thick at nearly the same radius. Unfortunately, this test is still sensitive to the large amount of scatter in the IGM density field (and in the lower-redshift Lyman-α forest that coincides with and hence obscures the Lyman-β measurement), so the current sample of < 10 quasars cannot distinguish R_b from $R_{\rm lim}$ – even though coincident Lyman-α and Lyman-β absorbers have been detected, it is not clear if they are due to a large swath of neutral IGM gas or a single absorber.[iv] Simulations suggest that increasing the sample of such quasars by a factor of a few could lead to useful constraints when $\bar{x}_{\rm HI} > 0.1$, the regime in which the finite bubble size starts to affect the Lyman-β near-zone size.

Another difficulty with near-zone measurements, just as with the classical prox-

[iv] These kinds of identifications are further complicated by the damping wing absorption that we will examine next.

imity effect, is the biased region in which the quasar lives. Although the gas is only significantly overdense in a relatively small region immediately around the quasar, even modest overdensities in the dark matter can lead to substantial overdensities in the biased galaxy population. Moreover, the ionized bubble generated by these galaxies reaches much larger distances than the galaxy overdensity itself – even the tens of comoving Mpc typical of a bright quasar's near-zone. The easiest way to understand this is to think of the overdense region as a piece of a Universe with $\Omega_m > 1$: in that case structure formation proceeds faster, because of the increased gravity, and both the local ionized fraction and the ionized bubbles themselves grow faster as well. This implies that the ionized fraction measured from the quasar near-zone will be biased relative to the true average.

10.4 THE RED DAMPING WING

If Lyman-α photons encounter nearly neutral gas with $\tau_\alpha > 10^5$, the broad Lyman-α absorption line can significantly affect their transfer through the IGM. Considering only the regime in which $|\nu - \nu_\alpha| \gg \Lambda_\alpha$ (and neglecting the broadening introduced by the finite temperature of the IGM), we may ignore the second term in the denominator of equation (4.8). If we assume that the IGM has a uniform neutral fraction $\bar{x}_D$ at all points between the edge of a source's local ionized bubble (which we call z_b) and $z_{\rm reion}$, this leads to an analytical result valid within the "red damping wing" of the Gunn-Peterson trough for the optical depth at an observed wavelength $\lambda_{\rm obs} = \lambda_\alpha(1+z)$:

$$\tau(z) = \tau_\alpha \bar{x}_D \left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)\left(\frac{1+z_b}{1+z}\right)^{3/2}\left[I\left(\frac{1+z_b}{1+z}\right) - I\left(\frac{1+z_{\rm reion}}{1+z}\right)\right], \tag{10.21}$$

for $z > z_b$, where

$$I(x) \equiv \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\ln\left[\frac{1+x^{1/2}}{1-x^{1/2}}\right] . \tag{10.22}$$

Note that here we define z as the redshift at which the observed photon would have passed through Lyman-α; however, when $z > z_b$ this never actually happens. This expression is only valid far from line center, but that is usually acceptable because the optical depth is so large there anyway. It also assumes $\Omega_m(z) \approx 1$, which is adequate at the high redshifts of interest $z \gg 1$.

At wavelengths for which $|x-1| \ll 1$, one can approximate the $I(x)$ factors with their asymptotic limits; in that case,

$$\tau(z) \approx \tau_\alpha \bar{x}_D \left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)\frac{c(1+z)}{H(z)}\left(\frac{1}{R_{b1}} - \frac{1}{R_e}\right), \tag{10.23}$$

where R_{b1} is the comoving distance to the edge of the source's ionized bubble and R_e is the comoving distance to the surface defining the "end" of reionization. As a rule of thumb, the damping wing optical depth approaches unity at a velocity offset of ~ 1500km s^{-1}, which corresponds to ~ 1 proper Mpc at $z \sim 10$.

The exciting prospect is that within this red damping wing, the optical depth experienced by the photons approaches order unity over a fairly wide range of redshifts: this means that the optical depth can be measured relatively easily, in contrast to the strongly saturated absorption at line center. Crudely, if we can therefore measure z_s and $\tau(z)$ we can obtain an estimate for the IGM neutral fraction.

Figure 10.4 illustrates the resulting absorption profiles for three choices of $\bar{x}_D = 0.9,\ 0.5,$ and 0.1 (thin dashed, solid, and dotted curves, respectively); in all cases we take $z_{\rm reion} \ll z_b$. Here the abscissa measures the wavelength offset from the source redshift z_s; we take z_b, where neutral gas first appears, to be 5 comoving Mpc from the source. Note that, especially for the more neutral cases, the absorption extends to quite large redshift offsets from line center: $z - z_s = 0.01(1 + z_s)$ translates to an observed wavelength offset of $12(1 + z_s)$ Å. The dot-dashed line shows the absorption profile of a single absorbing cloud at a fixed location (i.e., a DLA), normalized to have the same transmission at z_s as the $\bar{x}_D = 0.1$ curve.

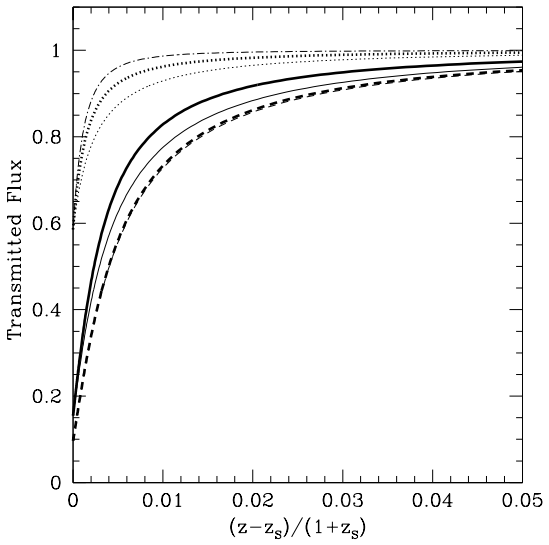

Figure 10.4 Damping wing absorption profiles, as a function of fractional wavelength offset from the source (at redshift z_s). The thick curves show the absorption profiles for $\bar{x}_D = 0.9,\ 0.5,$ and 0.1 assuming the "picket fence" model of absorption (with the dashed, solid, and dotted curves, respectively). Note that the two dashed curves overlap and are practically indistinguishable. The corresponding thin curves show the absorption profiles for uniformly ionized IGM normalized to the same transmission at z_s. The dot-dashed curve shows the profile of a DLA, normalized to the same transmission as the $\bar{x}_D = 0.1$ curves at z_s. Figure credit: Mesinger, A. & Furlanetto, S.R. 2008, MNRAS, 385, 1348.

Obviously, the IGM absorption profile is much gentler than that from a DLA, extending to much larger redshift offsets. Indeed, equation (10.23) shows that the optical depth scales as the inverse of the wavelength offset between the observed

wavelength and λ_α at the source redshift. In contrast, DLAs have $\tau \propto \Delta\lambda^{-2}$; the difference arises because the photon continues to redshift away from line center as it passes through the IGM, so a photon at a given wavelength experienced a larger optical depth than one would expect had it remained at a constant frequency through the entire column. In practice, this may be a crucial discriminant between absorption intrinsic to a high-redshift source (taking the form of a DLA) and that from the IGM. For example, nearly all GRBs at lower redshifts have associated high-column absorbers. The different absorption profiles are crucial for indentifying the nature of the Lyman-α absorption.

Unfortunately, the simple toy model we have used so far does not accurately describe the IGM during reionization, and the real absorption profiles are likely to be somewhere between these two limits. We have already seen that in most reionization scenarios the IGM has a two-phase structure, with seas of neutral gas surrounding bubbles of ionized matter. A typical line of sight through the IGM will therefore pass through a "picket fence" of absorbers composed of alternating patches of nearly neutral and nearly ionized gas. The resulting absorption profiles, shown for a toy model by the thick lines in Figure 10.4, are steeper than those in a uniform IGM (unless the ionized bubbles are very rare) but shallower than for a DLA: essentially, the photon passes through a series of DLAs separated by clear regions. Because their frequency still changes as they travel, they experience more absorption than for a single cloud.

Obviously, this introduces some significant complications into interpreting the damping wing. The easiest way to see this is to consider our crude estimate for the average ionized fraction in a uniform IGM from equation (10.23). Here we can estimate $\bar{x}_D$ from the absorption at a single wavelength, provided that we assume a $1/\Delta\lambda$ profile to be accurate. (Note that we could also estimate z_b from the peak of the absorption line.) In this "picket fence" model, the true optical depth is a sum over that from all the neutral stretches of the IGM, or

$$\tau(z) \approx \tau_\alpha \left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)(1+z)^2 \sum_i \left(\frac{1}{z - z_{b,i}} - \frac{1}{z - z_{e,i}}\right). \tag{10.24}$$

where the ith neutral patch stretches between $z_{b,i}$ and $z_{e,i}$. If we naively equate this true expression to equation (10.23) and solved for $\bar{x}_D$, we find

$$\bar{x}_D \approx (z - z_{b,1}) \left\langle \sum_i \left(\frac{1}{z - z_{b,i}} - \frac{1}{z - z_{e,i}}\right)\right\rangle. \tag{10.25}$$

If we take a particularly simple model for the picket fence absorbers, in which the ionized and neutral patches have fixed lengths R_b and fR_b, where $f = (1 - Q_{\rm HII})/Q_{\rm HII}$ ensures the proper filling fraction of the bubbles, we can perform this sum and calculate the bias in our estimator $\bar{x}_D$:

$$\bar{x}_D = \frac{1}{2}\sum_{k=1}^{\infty}\left[\frac{1}{(k-1/2)+(k-1)f} - \frac{1}{(k-1/2)+kf}\right] \tag{10.26}$$

$$= \pi(1 - Q_{\rm HII})\cot\left[\frac{\pi(1-Q_{\rm HII})}{2}\right]. \tag{10.27}$$

This difference $\bar{x}_D - (1 - Q_{\rm HII})$ is always positive and peaks at ~ 0.3 when $Q_{\rm HII} = 0.5$, though the fractional bias continues to increase as $Q_{\rm HII} \to 1$. The actual amount of the bias of course depends upon the particular model of reionization (and in particular the size distribution and clustering of the H II regions); more detailed simulations have comparable (though slightly smaller) bias. This means that the damping wing requires non-trivial modeling to interpret it properly in the context of reionization.

Even if this bias can be corrected, a second problem is that different lines of sight inevitably pass through different sets of ionized and neutral patches, so there can be large scatter in the absorption profiles even for a given $Q_{\rm HII}$ and bubble size distribution. This scatter becomes particularly important in the late stages of reionization, because the damping wing optical depth is rather sensitive to the size of the first neutral patch.

Figure 10.5 illustrates these two problems in the context of a more realistic semi-numeric model of reionization. The curves show the probability distribution of $\delta_{x_D} \equiv \bar{x}_D/(1 - Q_{\rm HII}) - 1$ for a variety of bubble filling factors. Note that the means of these distributions are non-zero (implying a bias in the estimator) and the scatter increases dramatically in the later stages of reionization. This means that reliable estimates of the IGM properties will require a large number of lines of sight with measured damping wings.

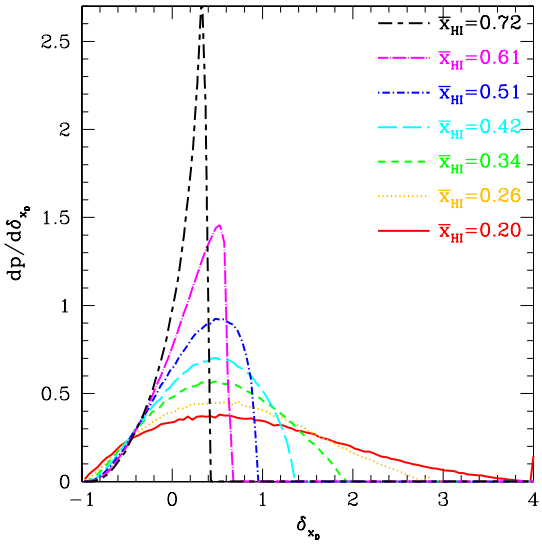

Figure 10.5 Probability distributions of the fractional bias in a simple damping wing estimate of the ionized bubble filling factor, $\delta_{x_D} \equiv \bar{x}_D/(1 - Q_{\rm HII}) - 1$. The different curves show different stages in reionization; all are computed with a semi-numeric simulation of reionization. Note that the mean is always non-zero, and the distribution becomes both wider and more biased as reionization progresses. Figure credit: Mesinger, A. & Furlanetto, S.R. 2008, MNRAS, 385, 1348.

Because the damping wing absorption profile must itself be measured, damping wing constraints on reionization require very bright sources. The two most likely candidates are quasars and GRBs. The former have the advantage of lying inside large H II regions, which decreases the bias and scatter in the estimators; however, they often also have substantial Lyman-α lines with unknown intrinsic properties, which complicates the measurement of the damping profile.

Figure 10.6 illustrates some of the complexities of a damping wing measurement with a quasar at $z = 7.085$, ULAS J1120+0641. We show the spectrum normalized to a composite spectrum constructed from lower-redshift quasars, which provides a surprisingly good fit to the data (though it appears to underestimate the Lyman-α line strength in this object). Provided that the template is accurate, the binned curve in the Figure therefore shows the transmission. As expected, this declines rapidly slightly blueward of the Lyman-α line center; this is the "near-zone" discussed above. This quasar has a very small near-zone, indicating either a high column density absorber along the line of sight, the presence of a substantially neutral IGM that the quasar still must ionize, or a very young age for the quasar. The smooth curves show the expected absorption for several IGM scenarios. The second curve from the top at the Lyman-α wavelength shows the absorption profile from a DLA 21 comoving Mpc in front of the source. The others show the absorption expected from a uniformly neutral IGM beginning 17.8 comoving Mpc in front of the quasar; these take $\bar{x}_D = 0.1, 0.5$, and 1, from top to bottom. Of these, the DLA profile appears to provide the best fit; however, more sophisticated fits taking into account the inhomogeneous ionization structure of the IGM could also match the data.

GRBs have much simpler intrinsic spectra (nearly power-law over this range), which makes extracting the damping wing easier. However, their host galaxies often have strong DLA absorbers, which interfere with the damping wing, and their position inside of small galaxies makes the bias and scatter large. It is not clear which will eventually prove more useful, though in either case constructing samples of many sources will be difficult. Also, in contrast to quasars, GRBs (and their faint host galaxies) have a negligible influence on the surrounding intergalactic medium. This is because the bright UV emission of a GRB lasts less than a day, compared with tens of millions of years for a quasar. Therefore, bright GRBs are unique in that they probe the true ionization state of the surrounding medium without modifying it.

10.4.1 Lyman-α Halos Around Distant Sources

As we have already discussed in the context of Lyman-α scattering within galaxies, Lyman-α line photons emitted by these galaxies are not destroyed but instead are absorbed and re-emitted as they scatter. For scattering in the uniform IGM, this problem is particularly simple and illuminates more of the physics of the scattering process.

Due to the Hubble expansion of the IGM around the source, the frequency of the photons is slightly shifted by the Doppler effect in each scattering event. As a result, the damping wing photons diffuse in frequency to the red side of the Lyman-α resonance. Eventually, when their net frequency redshift is sufficiently large,

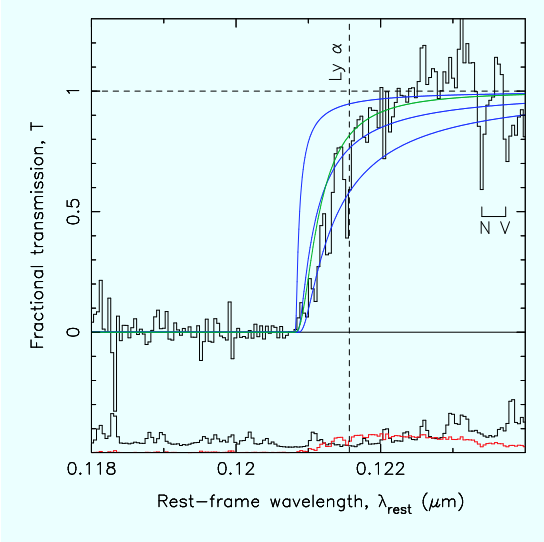

Figure 10.6 Rest-frame transmission profile of ULAS J1120+0641, in the region of the Lyman-α emission line, compared to several damping profiles. The transmission profile of ULAS J1120+0641, obtained by dividing the spectrum by a lower redshift quasar composite spectrum, is shown as the binned curve. The random error spectrum is plotted below the data. The other error curve shows the uncertainty in the Lyman-α equivalent width. Three of the four smooth curves in the upper panel show the expected absorption from an IGM damping wing with $\bar{x}_D = 1$, 0.5, 0.1 located a distance $R_b = 17.8$ Mpc in front of the quasar (bottom, second from bottom, and top curves at the Lyman-α wavelength). The other curve (second from top) shows a DLA absorber with $N_{\rm HI} = 4 \times 10^{20}$ cm^{-2} located a distance 21 Mpc in front of the quasar. Figure credit: Mortlock, D. et al., Nature, **474**, 616 (2011).

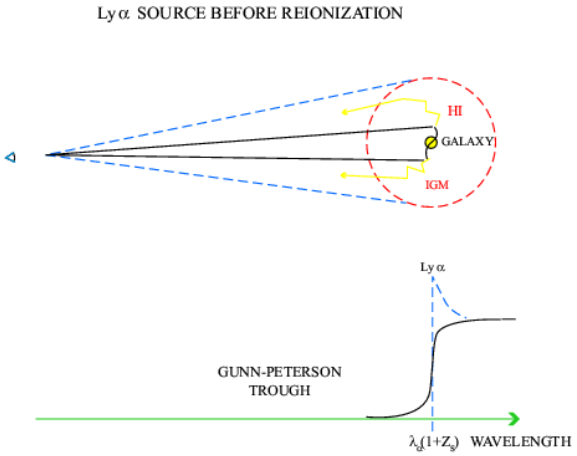

Figure 10.7 Halo of scattered Lyman-α line photons from a galaxy embedded in the neutral IGM prior to reionization (also called a *Loeb-Rybicki halo*). The line photons diffuse in frequency due to the Hubble expansion of the surrounding medium and eventually redshift out of resonance and escape to infinity. A distant observer sees a Lyman-α halo surrounding the source, along with a characteristically asymmetric line profile. The observed line should be broadened and redshifted by about one thousand $\mathrm{km\,s^{-1}}$ relative to other lines (such as Hα) emitted by the galaxy.

they escape and travel freely towards the observer (see Figure 10.7). As a result, the source creates a faint Lyman-α halo on the sky.[v] These *Loeb-Rybicki* Lyman-α halos can be simply characterized by the frequency redshift relative to the line center, $\nu_\star = |\nu - \nu_\alpha|$, which is required in order to make the optical depth from the source equal to unity. At high redshifts, the leading term in equation (10.21) yields

$$\nu_\star = 8.85 \times 10^{12}\ \mathrm{Hz} \times \left(\frac{\Omega_b h}{0.05\sqrt{\Omega_m}}\right)\left(\frac{1+z_s}{10}\right)^{3/2}, \qquad (10.28)$$

as the frequency interval over which the damping wing affects the source spectrum. A frequency shift of $\nu_\star = 8.85 \times 10^{12}$ Hz relative to the line center corresponds to a fractional shift of $(\nu_\star/\nu_\alpha) = (v/c) = 3.6 \times 10^{-3}$ or a Doppler velocity of $v \sim 10^3\ \mathrm{km\,s^{-1}}$. The Lyman-$\alpha$ halo size is then defined by the corresponding proper distance from the source at which the Hubble velocity provides a Doppler shift of this magnitude,

$$r_\star = 1.1\left(\frac{\Omega_b/0.05}{\Omega_m/0.3}\right)\ \mathrm{Mpc}. \qquad (10.29)$$

Typically, the observable Lyman-α halo of a source at $z_\mathrm{s} \sim 10$ occupies an angular radius of $\sim 15''$ on the sky (corresponding to $\sim 0.1 r_\star$) and yields an asymmetric

[v] The photons that begin blueward of Lyman-α and are absorbed in the Gunn-Peterson trough are also re-emitted by the IGM around the source. However, since these photons originate on the blue side of the Lyman-α resonance, they travel a longer distance from the source, compared to the Lyman-α line photons, before they escape to the observer. The Gunn-Peterson photons are therefore scattered from a larger and hence dimmer halo around the source.

line profile as shown in Figures 10.7 and 10.8. The scattered photons are highly polarized and so the shape of the halo would be different if viewed through a polarization filter.

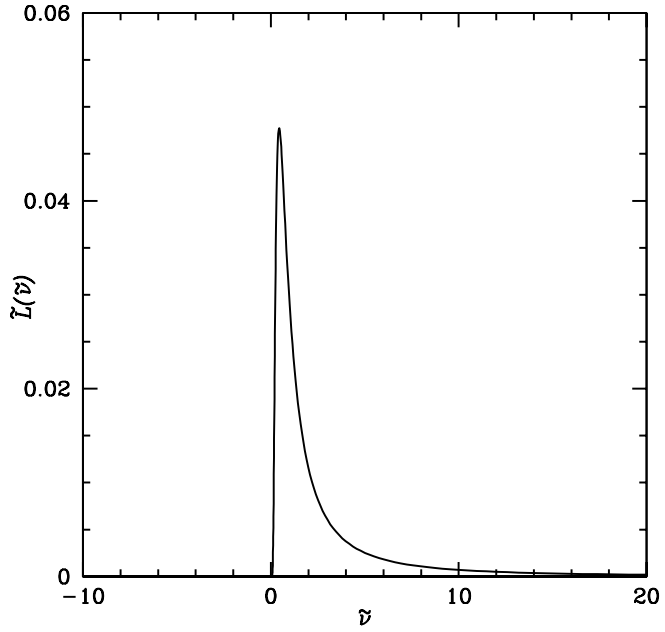

Figure 10.8 Monochromatic photon luminosity of a Lyman-α halo as a function of normalized frequency shift from the Lyman-α resonance, $\tilde{\nu} \equiv (\nu_\alpha - \nu)/\nu_\star$. Note that only the photons inside the red damping wing scatter in this compact halo; those on the blue side of Lyman-α scatter at much larger distances. The observed spectral flux of photons $F(\nu)$ (in photons cm^{-2} s^{-1} Hz^{-1}) from the entire Lyman-α halo is $F(\nu) = (\tilde{L}(\tilde{\nu})/4\pi d_{\rm L}^2)(\dot{N}_\alpha/\nu_\star)(1+z_s)^2$ where $\dot{N}_\alpha$ is the production rate of Lyman-α photons by the source (in $\rm photons\,s^{-1}$), $\nu = \tilde{\nu}\nu_\star/(1+z_{\rm s})$, and $d_{\rm L}$ is the luminosity distance to the source. Figure credit: Loeb, A. & Rybicki, G. B. *Astrophys. J.* **524**, 527 (1999); see also **520**, L79 (1999)].

Detection of the diffuse Lyman-α halos around bright high-redshift sources (which are sufficiently rare so that their halos do not overlap) would provide a unique tool for probing the distribution and the velocity field of the neutral IGM before the epoch of reionization. The Lyman-α sources serve as lamp posts which illuminate the surrounding H I fog. However, due to their low surface brightness, the detection of Lyman-α halos through a narrow-band filter is much more challenging than direct observation of their sources. Moreover, the velocity fields around these galaxies may be complicated by winds and infall, which would affect the line brightness and profile in similar ways to those discussed in §10.1.1.

10.5 THE Lyman-α FOREST AS A PROBE OF THE REIONIZATION TOPOLOGY?

Given the utility of the Lyman-α forest for understanding the ionization state of the IGM at low and moderate redshifts, extension of these techniques to the cosmic dawn is an obvious test of the topology and nature of the reionization process. However, we have already seen that the Gunn-Peterson optical depth is large at this time, even in highly ionized gas. Thus, we should not expect a clear signature of the ionized bubbles.

Nevertheless, the nature of the transformation from a bubble-dominated IGM to the post-reionization "web-dominated" IGM does offer some hope. Once the ionized bubbles become larger than the mean free path of the ionizing photons, the ionizing background saturates – even if the Universe were fully-ionized, the metagalactic background would not increase. Thus, in bubbles that have reached this saturation limit, we can expect nearly as much transmission as in the post-reionization IGM.

The key difference is the presence of the damping wing from the neutral gas surrounding each ionized bubble. With the rule of thumb that $\tau_d < 1$ only at distances > 1 proper Mpc from fully neutral gas, this requires that ionized bubbles be at least a few proper Mpc large in order to allow for any transmission. Fortunately, in most reionization models this constraint is easily fulfilled, at least in the latter half of reionization (see Fig. 8.3, for example).

On the other hand, even a moderate damping wing optical depth still increases the required transmission allowed by the residual neutral gas inside the bubble. Because bubbles that allow transmission must be very large, and thus contain an enormous number of luminous sources, their ionizing background is quite uniform (except at the edges of the bubble, but there the damping wing is large anyway). Thus, just as in the post-reionization IGM, transmission will come from highly-underdense voids in which the neutral fraction is small. Equation (10.15) shows that $\tau < 1$ requires $\delta < 0.1$–0.2. Such deep voids are very rare at high redshifts, because structure formation is still in its infancy – and of course such regions are largely empty of galaxies and so are likely to remain neutral throughout nearly all of reionization.

Thus, we expect transmission spikes to be extremely rare during reionization, but not impossible. With models for the H II region sizes, the emissivity of the galaxies driving reionization, and of the density distribution of the IGM, it is not difficult to estimate the possible abundance of transmission features. Figure 10.9 shows an optimistic example calculation for transmission at $z = 6.1$ (in the range probed by the highest-redshift known quasars). In this case, the IGM density distribution is calibrated to numerical simulations at $z = 2$–4. The curves show that observable transmission gaps with $\tau < 2.3$ occur only about once per $\Delta z \sim 3$.

In reality, transmission will be even more rare because this simple calculation makes the optimistic assumption that photons travel to the edge of their bubble, without any limits from LLSs in the IGM. But even so, Figure 10.9 shows that they are sufficiently rare that precise quantitative constraints on reionization from the

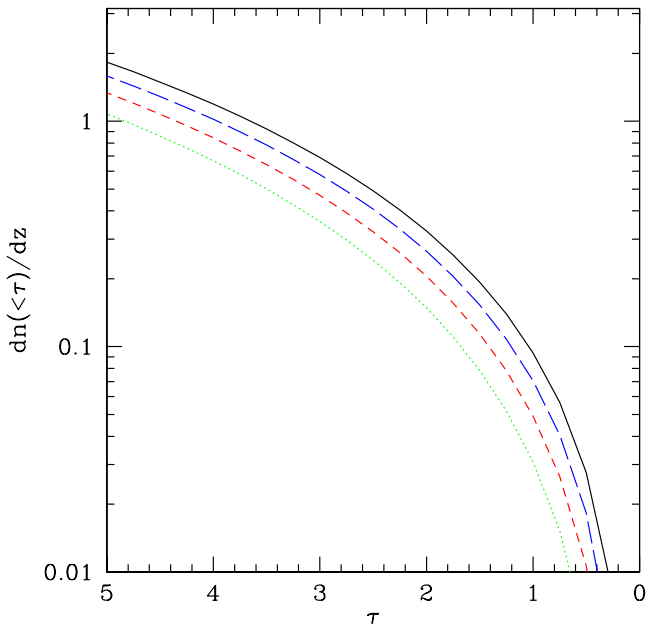

Figure 10.9 A model for the expected cumulative number of transmission features at $z = 6.1$ if the IGM has $Q_{\rm HII} = 0.9,\ 0.85,\ 0.8$, and 0.75 (solid, long-dashed, short-dashed, and dotted curves, respectively). The model uses the excursion set model for reionization (see §8.5) and an inhomogeneous IGM density distribution calibrated to simulations at lower redshifts. Figure credit: Furlanetto, S.R. et al. 2004, MNRAS, 354, 695.

gaps will require much larger samples of quasars or GRBs than currently available. Drawing conclusions about reionization from the forest is instead very difficult. Indeed, some simulations of the reionization process show that the present data cannot even rule out reionization ending at $z < 6$, since some small pockets of neutral gas could remain, buried inside the long stretches of saturated absorption that are common at this time.

Instead, the Lyman-α forest (especially together with absorption in Lyman-β and Lyman-γ) is best at constraining the very end of the reionization era, as discussed in §4.6, unless the red damping wing can be measured on its own.

10.6 LYMAN-α EMITTERS DURING THE REIONIZATION ERA

We now return to discuss the properties of more normal galaxies that have Lyman-α lines, commonly referred to as Lyman-α emitters or LAEs. We saw in §9.2.1 that this strong emission line provides a convenient marker for young star-forming galaxies, and one of the most efficient ways to find distant galaxies is with narrow-

band searches that identify sources with strong emission lines in a narrow redshift range.

We have seen in §10.1.1 that the intrinsic properties of the Lyman-α line depend on a host of complex factors. However, we have also found in §10.3.1 and 10.4 that resonant absorption in the ionized IGM and much stronger absorption from neutral gas – even from the damping wing once the photon has passed through resonance – can also strongly affect the line shape. These latter effects make the Lyman-α emission lines of galaxies an interesting and potentially powerful probe of IGM properties. However, we must always bear in mind the complexity of the intrinsic line profile as an important source of systematic confusion for such a probe.

Figure 10.10 shows how this IGM reprocessing can dramatically alter the observed line intensity and profile; the top panel shows the lines, while the bottom panel shows the corresponding optical depth profiles. In the top panel, the upper dotted curve shows the assumed intrinsic line, which we place at $z = 10$ and take as a Gaussian with width $27\,\mathrm{km\ s^{-1}}$ (these are arbitrary choices chosen for illustrative purposes). The other curves show the effects of IGM reprocessing, including both the damping wing from fully neutral gas at a distance R_b from the line source (with R_b decreasing from top to bottom) and resonant scattering from the ionized medium within (except for the lower dotted curve). The optical depths providing this absorption are shown in the bottom panel: the nearly-horizontal lines are the damping wing optical depths (with R_b increasing from bottom to top), while the dotted curve shows the resonant value.

Note that the resonant absorption is large everywhere blueward of line center, but it is modest or negligible on the red side. This is a rather generic result (here we have included only the ionization from the galaxy itself, which dominates on the relevant scales, so the ionization structure on large scales is negligible); in general, we expect LAEs at $z > 5$ to have asymmetric line profiles, with the blue side cut off by resonant IGM absorption.

However, the damping wing absorption that affects the red side (as well as the blue side) depends sensitively on the large scale environment, and in particular the displacement from the source to the nearest neutral gas. We see here that a bubble with $R_b = 1$ *proper* Mpc provides $\tau_D \approx 1$; in fact, this rule of thumb works reasonably well throughout the relevant high-z regime.

We therefore expect that as we penetrate farther back into the reionization era, with the bubbles growing smaller and smaller, more and more of their Lyman-α lines will be extinguished by the neutral gas. In the remainder of this section we will explore the consequences of this expectation for LAE surveys during reionization.

10.6.1 Galaxies within Ionized Bubbles

In order to understand the interplay between the damping wing and galaxy populations, we must first understand how galaxies populate the H II regions that surround them. Fortunately, because we can use the same methods – the excursion set formalism – to compute the halo and ionized bubble abundances, this is a relatively easy task.

Consider an ionized bubble with mass m_b and a mean overdensity δ_b; according

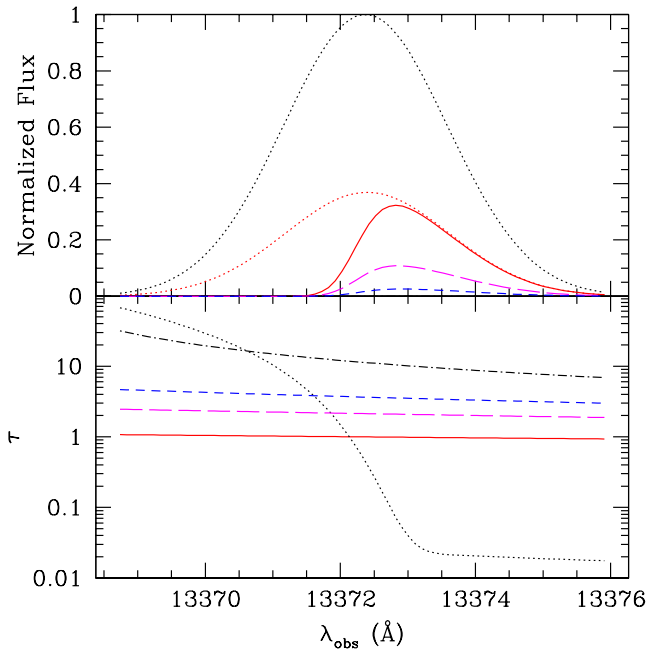

Figure 10.10 *Top:* Example line profiles for a galaxy at $z = 10$. The upper dotted curve shows the intrinsic line profile, assumed to be a Gaussian with standard deviation 27 km s^{-1}. The solid, long-dashed, and short-dashed curves show the observed line after reprocessing through the IGM; they place the galaxy in ionized bubbles with radii $R_b = 10$, 5, and 3 comoving Mpc, respectively. The lower dotted curve shows the line if we neglect resonant absorption within the ionized bubble, assuming $R_b = 10$ Mpc. *Bottom:* The dotted line shows the resonant absorption from the ionized bubble. The solid, long-dashed, short-dashed, and dot-dashed curves show the damping wing optical depth for $R_b = 10$, 5, 3, and 1 Mpc, respectively. Figure credit: Furlanetto, S.R. et al. 2004, MNRAS, 354, 695.

to the model in §8.5, this overdensity is exactly that required for a collapse fraction large enough to produce one ionizing photon per hydrogen atom inside the bubble, so $\delta_b = B(m_b)$. We wish to know the abundance of galaxies as a function of mass m within this ionized bubble, $n(m|m_b)$.

In the excursion set picture (see §3.5.2), this is simply proportional to the fraction of random walks that begin at (m_b, δ_b) and end at (m, δ_c), where $\delta_{\rm crit}$ is the critical linearized overdensity for halo collapse (which is a function of m in, for example, the Sheth-Tormen model). But this problem is actually identical to the "extended Press-Schechter" problem, in which we calculated the progenitors of a given halo at an earlier redshift: the only difference is that here our "halo" is a bubble and we work at the same redshift – which is possible because the criterion for an ionized bubble requires a lower overdensity than halo collapse itself.

Thus we can immediately write

$$n(m|m_b) = \sqrt{\frac{2}{\pi}} \frac{\bar{\rho}}{m^2} \left| \frac{d\ln\sigma}{d\ln m} \right| \frac{\sigma^2[\delta_{\rm crit}(z) - B(m_b)]}{(\sigma^2 - \sigma_b^2)^{3/2}} \exp\left\{ -\frac{[\delta_{\rm crit}(z) - B(m_b)]^2}{2[\sigma^2 - \sigma_b^2]} \right\}, \tag{10.30}$$

where $\sigma^2 = \sigma^2(m)$ and $\sigma_b^2 = \sigma^2(m_b)$.

We can also perform the reverse calculation (analogous to the distribution of halo descendants) to compute the probability $p_b(m_b|m)$ that a halo of mass m is part of a bubble of mass m_b. Figure 10.11 shows the results of this calculation for a small halo ($m_h = 10^9\ M_\odot$) and a moderately large one ($m_h = 10^{11}\ M_\odot$). The different curves in each panel correspond to a sequence of ionized fractions in a model of reionization. Unsurprisingly, the median bubble size increases as reionization progresses (because all bubbles grow with time), but note that it also strongly depends on the halo mass: large galaxies are far more likely to reside in large bubbles than average galaxies. This is just another manifestation of the increasing bias of galaxies with their mass.

10.6.2 LAE Number Counts During Reionization

Next let us imagine performing a sequence of narrowband Lyman-α searches at progressively larger redshifts. We expect that, once the typical bubble size falls below ~ 1 proper Mpc, the IGM damping wing will also start to extinguish the lines even if the galaxies still exist. We might therefore imagine a simple counting exercise as a test for reionization, aiming to see a decline in the abundance of LAEs.

Of course, there are many other reasons why the LAE density may decline – most obviously, the halo mass function changes rapidly with z at these early times, so the galaxy abundance most likely does as well. Ideally one would therefore calibrate it to a broadband galaxy survey that is not subject to the same selection effects – if the LAE abundance declines precipitously while the overall galaxy density declines only gently, that would be good evidence for IGM absorption. Note, however, that the complicated physics of Lyman-α generation and transfer within galaxies always leaves some room for doubt, since such a decline could also be attributed to the evolving IMF of stars or changes in their dust content.

Nevertheless, this simple test is very attractive. We can use the excursion set formalism described in §10.6.1 to estimate how the abundance would decline. We ignore the effects of resonant absorption (since they depend on the local environment of the galaxy and hence are unlikely to evolve rapidly during reionization) but include the damping wing absorption from neutral gas in the IGM. Let us suppose that the survey is sensitive to sources with $L > L_{\rm min}$. If we then take $L \propto m$ for simplicity, a galaxy halo of mass m will be detected only if the damping wing has $\tau_D < \ln(m/m_{\rm min})$, where $L(m_{\rm min}) = L_{\rm min}$. Then the number density of observable galaxies is

$$n(>L) = \int dm_b\, n_b(m_b) V_b \int_{m_D}^{\infty} dm\, n(m|m_b), \tag{10.31}$$

where m_D is the minimum halo mass that remains observable inside a bubble of mass m_b and volume V_b. Note that m_D *decreases* with m_b, since larger bubbles

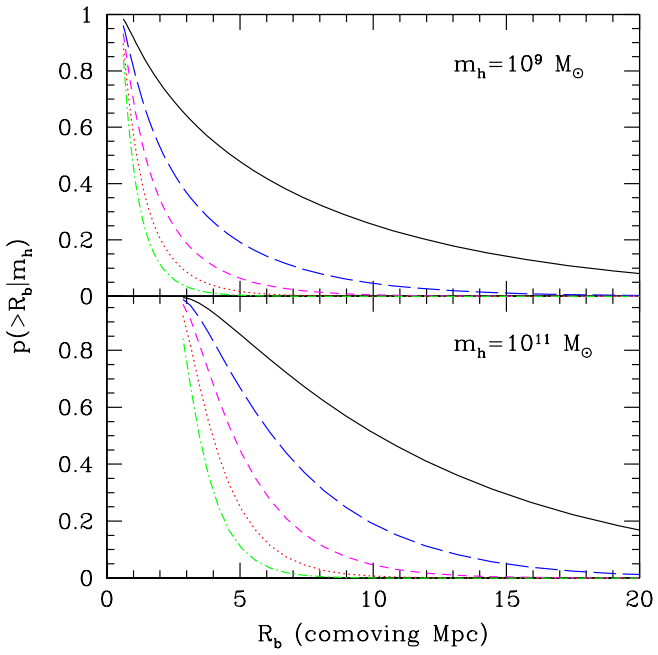

Figure 10.11 Probability that halos with $m_h = 10^9$ and $10^{11}\ M_\odot$reside in ionized bubbles larger than a given radius R_b. Here we use the excursion set model of reionization with $\zeta = 40$; the bubble sizes are relatively independent of this choice, for a fixed $Q_{\rm HII}$, but the halo populations themselves are highly-redshift dependent. In each panel, the curves correspond to $z = 12$ ($Q_{\rm HII} = 0.74$, solid), $z = 13$ ($Q_{\rm HII} = 0.48$, long dashed), $z = 14$ ($Q_{\rm HII} = 0.3$, short-dashed), $z = 15$ ($Q_{\rm HII} = 0.19$, dotted), $z = 16$ ($Q_{\rm HII} = 0.11$, dot-dashed), Figure credit: Furlanetto, S.R. et al. 2004, MNRAS, 354, 695.

cause less damping wing absorption. Of course, in reality τ_D is a function not only of bubble size but of a galaxy's position within the bubble: those at the edge always experience strong absorption.

Nevertheless, this simple model is in good agreement with more detailed calculations using simulations of reionization (either full-scale or semi-numerical). Figure 10.12 shows the luminosity function at several different neutral fractions (including fully ionized, top curve) measured in a semi-numerical simulation. Clearly damping wing absorption from the neutral gas can have an enormous effect on the observed abundance of galaxies in these surveys.

The detailed calculation reveals two interesting effects. First, the fractional decline is relatively modest (no more than a factor ~ 2) until $Q_{\rm HII} < 0.5$; beyond that point the abundance declines precipitously. This is because the ionized bubbles have characteristic sizes ~ 10 comoving Mpc, or ~ 1 proper Mpc, when $Q_{\rm HII} \sim 0.5$. Larger bubbles, late in reionization, have $\tau_D < 1$ and so have only a small effect on the observed abundance.

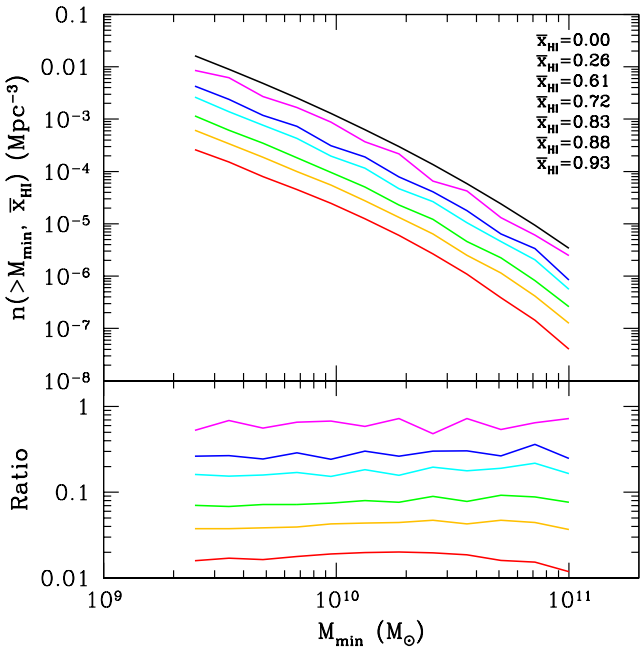

Figure 10.12 Luminosity function of LAEs at $z = 9$ in a semi-numeric simulation of reionization, as a function of the mean neutral fraction $\bar{x}_{\rm HI}$. The sequence of curves from top to bottom goes from small to large $\bar{x}_{\rm HI}$. The bottom panel shows the ratio of the curves to that in a fully-ionized Universe. Figure credit: Mesinger, A. & Furlanetto, S.R. 2008, MNRAS, 386, 1990.

The second factor is visible in the bottom panel of Figure 10.12: evidently the fractional decline in LAE abundance is nearly independent of halo mass (or intrinsic luminosity). This occurs because the distribution ofτ_D is quite broad (roughly lognormal), due not only to the range of halo sizes but also to the distribution of galaxies within each bubble. For faint galaxies, which roughly follow a power-law intrinsic distribution, the convolution of these two effects preserves the power law. At the bright end, where the intrinsic luminosity function declines exponentially, the breadth of the τ_D distribution masks the change in slope.

10.6.3 LAE Clustering During Reionization

The fact that galaxies within large ionized bubbles remain (relatively) unattenuated while those inside of small bubbles will be extinguished by the damping wing suggests that not just the mean number density of LAEs will change throughout reionization, but that their spatial distribution will evolve as well. Figure 10.13 shows this explicitly. Each panel shows a slice through a semi-numeric simulation of reionization; here we fix $z = 9$ and vary the ionized fraction across the panels (from fully-ionized at left to $\bar{x}_{\rm HI} = 0.77$ at right). Each white dot corresponds to

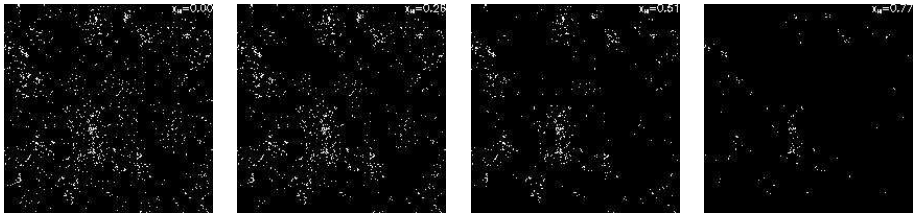

Figure 10.13 Maps of visible LAEs at $z = 9$ in a semi-numeric simulation, assuming $\bar{x}_{\rm HI} \approx$ 0, 0.26, 0.51, 0.77, from left to right. All slices are 250 Mpc on a side and 20 Mpc deep. We assume that all halos with observed luminosities greater than that corresponding to an unattenuated galaxy with $M > 1.67 \times 10^{10}\ M_\odot$ are visible. Figure credit: Mesinger, A. & Furlanetto, S.R. 2008, MNRAS, 386, 1990.

a galaxy with an observable Lyman-α line, assuming the same model as the last section for their luminosity function. The overall trend is clear: galaxies that are relatively isolated in the left-most panel disappear first, while those that are part of a strong overdensity (near the bottom center of the image) remain visible even to large neutral fractions.

The best way to describe this phenomenon quantitatively is through the clustering of the galaxies. A simple toy model illustrates how it enhances the apparent clustering on small scales (relative to galaxies observed in the continuum, for example). Suppose that galaxies with number density $\bar{n}$ are distributed randomly throughout the universe but that we can only observe those with at least one neighbor within a sphere of volume $V \ll \bar{n}^{-1}$. Assuming a Poisson distribution, the number density of observed objects would be

$$n_{\rm obs} = \bar{n}(1 - e^{-\bar{n}V}). \tag{10.32}$$

As usual the correlation function of the observed sample is defined through the total probability of finding two galaxies in volumes δV_1 and δV_2,

$$\delta P = n_{\rm obs}^2\,(1 + \xi)\,\delta V_1\,\delta V_2. \tag{10.33}$$

However, we know that every observed galaxy has a neighbour within V; thus

$$\delta P = n_{\rm obs}\,\delta V_1\,(\delta V_2/V) \tag{10.34}$$

for small separations (where the factor $\delta V_2/V$ assumes the neighbor to be randomly located within V). Thus,

$$\xi = 1/(n_{\rm obs}V) - 1 \tag{10.35}$$

on such scales: even though the underlying distribution is random, the selection criterion induces clustering. Note that it can be extremely large if $V \ll n_{\rm obs}^{-1}$.

On large scales, the modulation takes a different form. An observed galaxy resides in a large bubble, corresponding to an overdense region. Because of the bias of the underlying dark matter field, that overdense region will tend to lie near other overdense regions – and hence other large bubbles. Thus, we will be more likely to see galaxies near the original object than in an average slice of the universe.

Because we do not see similar galaxies in small (less-biased) bubbles, the large-scale bias will generically be larger than that intrinsic to the galaxies.

Because these two effects have different amplitudes, the bubbles introduce a scale-dependent bias to the correlation function of galaxies, with a break at $r \approx R_c$, where R_c is the characteristic size of the ionized bubbles. Again using the excursion set formalism, we can estimate this modified bias in the limits $r \ll R_c$ and $r \gg R_c$.

By analogy with the halo model for the density field, these limiting regimes correspond to correlations between galaxies within a single bubble and within two separate bubbles. We begin with large scales: the observed clustering is the average bias of the bubbles weighted by the number of galaxies in each H II region (analogous to the two-halo term for the density field):

$$b_{r=\infty} = \int dm_b \, n_b(m_b) \, b_b(m_b) \, V_b \int_{m_D}^{\infty} dm_h \, \frac{n_h(m_h|m_b)}{\bar{n}_{\rm gal}}, \tag{10.36}$$

where we integrate only over those haloes visible after damping wing absorption and $\bar{n}_{\rm gal}$ is the mean number density of observable galaxies. Following the procedure outlined in §3.6, we can estimate the bias b_b of H II regions as[vi]

$$b_b(m_b) = 1 + \frac{B(m_b)/\sigma^2(m_b) - 1/B_0(m_b)}{D(z)}. \tag{10.37}$$

(Note that unlike the halo bias we can have $b_b < 0$: late in reionization, small bubbles are truly *anti*-biased because dense regions have already been incorporated into large ionized regions.)

The behavior on small scales is somewhat more subtle. If galaxies were randomly distributed within each bubble, the simple argument in the first paragraph of this section suggests that the correlation function would just be the weighted average of the number of pairs per H II region. However, in addition to the increase in the number of galaxies in each bubble, the galaxies also trace density fluctuations within each bubble. On moderately small scales were nonlinear evolution in the density field may be neglected, we therefore write

$$b_{\rm sm}^2 = \int dm_b \, n_b(m_b) \, V_b \, b_h^2(m_b) \, \frac{\langle N_{\rm gal}(N_{\rm gal} - 1)|m_b\rangle}{\bar{N}_{\rm gal}^2}, \tag{10.38}$$

where $\bar{N}_{\rm gal} = \bar{n}_{\rm gal} V_b$, $\langle N_{\rm gal}(N_{\rm gal} - 1)|m_b\rangle$ is the expected number of galaxy pairs within each bubble, and b_h^2 measures the excess bias of these haloes inside each bubble. Note the similarity to the halo-model calculation of the galaxy power spectrum here; in fact this form can be derived formally by constructing the galaxy density field from bubbles and their constituent haloes, in analogy to the halo model. This term then corresponds to the "two-halo, one-bubble" term in such a treatment; i.e., correlations between two particles that lie in the same bubble but different dark matter haloes. The "bubble profile" describing the distribution of galaxies within the bubble turns out to be proportional to the square root of the linear matter correlation function. Provided that the typical bubbles have more than two galaxies, we

[vi] This equation does not work late in reionization, because the physical requirement that $Q_{\rm HII} \leq 1$ caps the effective number density and hence the bias. In this regime numerical simulations are necessary; fortunately this regime is also the least interesting from the viewpoint of clustering.

can write the expected number of pairs as

$$\langle N_{\rm gal}(N_{\rm gal}-1)|m_b\rangle \approx \max\{0, \bar{N}_{\rm gal}(m_b)[\bar{N}_{\rm gal}(m_b)-1]\}. \quad (10.39)$$

The remaining factor is $b_h(m_b)$. It may seem reasonable to take this to be the mean value of the usual excursion set halo bias, evaluated over $n(m_h|m_b)$. However, the pair density inside each bubble *already* includes much of this bias because it counts the number of galaxies in a region with overdensity $\delta_b = B$. We therefore only want the "excess" bias of the galaxies relative to density fluctuations on scales smaller than m_b, which is the bias evaluated from the conditional mass function in equation (10.30). Following the excursion set definition of this bias, we have

$$b_h(m_h|m_b) = 1 + \frac{(\delta_c - \delta_x)^2/(\sigma^2 - \sigma_b^2) - 1}{\delta_c(z=0) - \delta_x(z=0)}. \quad (10.40)$$

We show the resulting bias at $z = 10$ as a function of $Q_{\rm HII}$ in Figure 10.14. In each panel, the different curves take different galaxy populations, with smaller galaxies having less net bias. Panels *(a)* and *(b)* show $b_{\rm sm}$ and $b_{\rm r=\infty}$. We scale the results to the bias $\bar{b}_h$ intrinsic to the galaxy population if absorption could be ignored. Panel *(c)* shows the ratio $b_{r=\infty}/b_{\rm sm}$, illustrating the magnitude of the "break" in the linear bias. We emphasize that the scale at which the break occurs will evolve throughout reionization along with the characteristic bubble size R_c; for illustrative purposes we mark several values of R_c.

Clearly, both $b_{\rm sm}$ and $b_{r=\infty}$ decrease throughout reionization. The large-scale bias decreases because the ionized regions must lie nearer to the mean density (and hence be less biased) as $Q_{\rm HII} \to 1$: this behaviour must be generic to any model in which reionization begins in overdense regions. The small-scale bias decreases because bubbles large enough to allow transmission become common: early on, only those galaxies with near neighbors are visible, so the correlations are strong. In the middle and final stages of reionization, most galaxies lie inside bubbles large enough to permit transmission, so more typical galaxies become visible and $b_{\rm sm} \to \bar{b}_h$.

These qualitative results also hold true in more detailed calculations with numerical simulations. Figure 10.15 shows the estimated angular correlation function (i.e., the three-dimensional correlation function projected on the plane of the sky) from a radiative transfer simulation of LAEs at $z = 6.6$, the highest redshift window easily visible to a ground-based telescope. The different curves in each panel correspond to different ionized fractions; the different panels describe different surveys, with the top panel comparable to existing capabilities and the others a few times larger. Note the enhancement in small scale correlations at small ionized fractions; this is the same effect we have described with $b_{\rm sm}$. The large-scale power is also enhanced, but it is much less sensitive to $\bar{x}_i$.

Although the correlation function and power spectrum (and through them the linear bias) are the most straightforward manifestations of the increased clustering, the "mask" applied to the galaxy distribution is itself non-Gaussian, so other clustering statistics – such as counts-in-cells or higher-order correlations – are also useful. All of these probes follow the qualitative behavior of the bias, increasing most dramatically early in the reionization process.

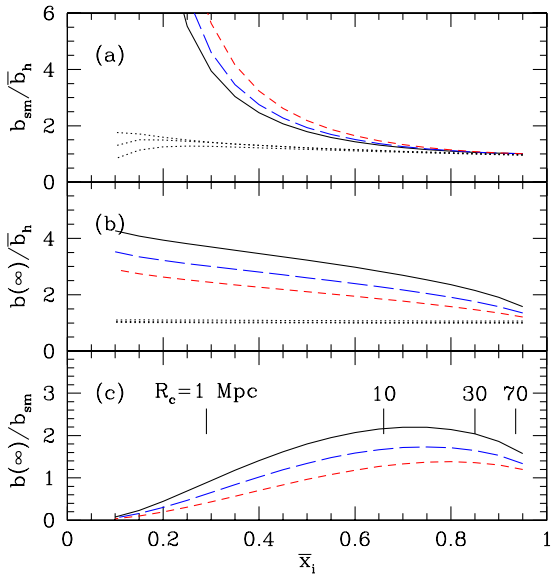

Figure 10.14 *(a)*: Predicted small-scale bias of LAEs at $z = 10$, relative to the bias expected if all galaxies above the mass threshold were visible. This applies to separations larger than the nonlinear scale but smaller than the characteristic bubble size R_c. The solid, long-dashed, and short-dashed curves take $m_{\rm obs,min} = 10^8$, 10^9, and $10^{10} M_\odot$, respectively. The dotted curves show the predicted galaxy bias, neglecting absorption, relative to its true value (the small errors at early times result from the approximations described in the text). *(b)*: Predicted large-scale bias at $z = 10$, relative to the bias expected if all galaxies above the mass threshold were visible. *(c)*: Ratio of large to small scale bias; the transition between the two regimes will occur roughly at R_c, which is marked for a few different values of the ionized fraction $\bar{x}_i$. Figure credit: Furlanetto, S.R. et al. 2006, MNRAS, 365, 1012.

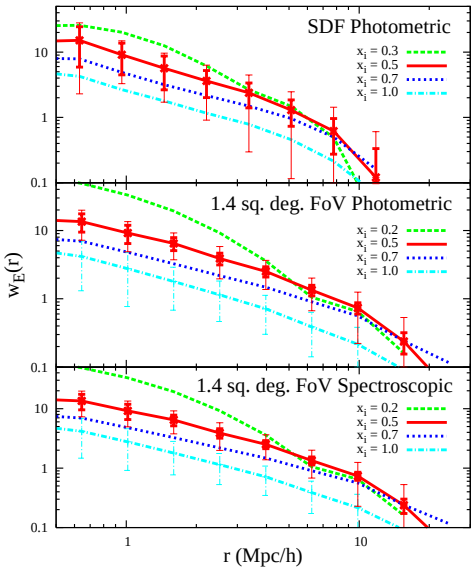

Figure 10.15 Angular correlation function of LAEs in a radiative transfer simulation of reionization. The simulation takes $z = 6.6$ and assumes all LAEs with an observed luminosity greater than the intrinsic luminosity of a halo with $m = 7 \times 10^{10}\ M_{\odot}$are visible. The different curves in each panel assume different ionized fractions. The top panel estimates the errors for existing surveys with the Subaru Deep Field in which LAEs are detected photometrically. The other two panels assume larger surveys (with ~ 5 times more LAEs); the middle panel assumes a photometric survey, while the bottom one assumes the LAEs can be selected spectroscopically. In each one, the thick error bars include Poisson fluctuations in the galaxy counts, while the thin curves also include cosmic variance. Figure credit: McQuinn, M. et al. 2007, MNRAS, 381, 75.

Both the analytic and numeric approaches show that the bias increases with redshift by a large factor, at least doubling and sometimes increasing by an even larger amount, especially on large scales. This, together with the change in the shape of the LAE correlation function with respect to the dark matter, makes the clustering signature much more robust to uncertainties in the nature of the LAE hosts. This is because the linear bias is a relatively slowly-varying function of halo mass and redshift; mimicking the shift due to reionization would require a drastic change in the properties of the galaxies.

However, it is worth emphasizing again that the radiative transfer of Lyman-α photons through the IGM is a complex process, and it can affect the observed clustering even after reionization is complete (thus the resonant absorption, which we have neglected in this section, can also be important). Interestingly, the frequency dependence of the scattering process induces anisotropies, generating clustering signatures analogous to redshift-space distortions. Fortunately, this component should not evolve as rapidly during reionization as the damping wing.

10.6.4 Lyman-α Blobs

A particularly interesting example of Lyman-α line emission in the interface between galaxies and the IGM are the so-called "Lyman-α blobs" (LABs) originally discovered in narrowband images at moderate redshifts ($z \sim 3$). So far, several tens of LABs have been found in the redshift range $z \sim 2$–7, making them much more common than initially expected. These blobs have a range of properties, but all are characterized by significantly extended Lyman-α line emission (ranging in size from ~ 10 kpc "halos" around star-forming galaxies to > 150 kpc giants with no obvious central galaxy in the rest-frame ultraviolet). Some appear to be diffuse elliptical objects, while others are much more filamentary. The brighter objects, with line luminosities $L > 10^{44}$erg s^{-1}, are extraordinarily powerful, corresponding to star formation rates $> 50 M_{\odot}$ yr^{-1}. The lines can be quite broad but do not show any unusual features like double-peaked profiles. Two example objects are shown in Figure **??**.

Bright LABs are typically located near massive galaxies that reside in dense regions of the Universe. Multi-wavelength studies of LABs reveal a clear association of the brighter blobs with sub-millimeter and infrared sources which form stars at exceptional rates of $\sim 10^3 M_{\odot}$ yr^{-1}, or with obscured active galactic nuclei (in fact, strong Lyman-α emission has been known for many years to surround some high-redshift radio galaxies). However, other blobs have been found that are not associated with any source powerful enough to explain the observed Lyman-α luminosities.

The origin of LABs is still unclear. Some models relate LABs to cooling radiation from gas assembling into the cores of galaxies. Other models invoke photoionization of cold ($T \sim 10^4$ K), dense, spatially extended gas by an obscured quasars or extended X-ray emission; the compression of ambient gas by superwinds to a dense Lyman-α emitting shell; or star formation triggered by relativistic jets from AGN. The latest models relate LABs to filamentary flows of cold ($\sim 10^4$K) gas into galaxies, which are generically found in numerical simulations of galaxy for-

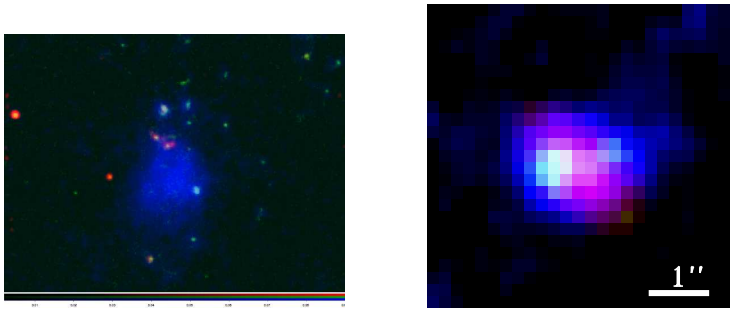

Figure 10.16 *Left:* A false color image of a Lyman-α blob (LAB) at a redshift $z = 2.656$. The hydrogen Lyman-α emission is shown in blue, and images in the optical V-band and the near-infrared J and H bands are shown in green and red, respectively. Note the compact galaxies lying near the northern (top) end of the LAB. The Lyman-α image was obtained using the SuprimeCam imaging camera on the Subaru Telescope, and the V, J, and H band images were obtained using the ACS and NICMOS cameras on the Hubble Space Telescope. This LAB was originally discovered by the Spitzer Space Telescope. Image credit: Prescott, M., & Dey, A. (2010). *Bottom:* A false color image of an LAB at a redshift $z = 6.6$, obtained from a combination of images at different infrared wavelengths. Image credit: Ouchi, M. et al. *Astrophys. J.* **696**, 1164 (2009).

mation. These cold flows contain $\sim$ 5–15% of the total gas content in halos as massive as $M_{\rm halo} \sim 10^{12}$–$10^{13} M_\odot$.

Although these objects have only been observed in detail so far at low redshifts, similar mechanisms offer the prospect of learning not only about star formation inside of high-redshift galaxies and the gross properties of the IGM but also about the detailed structure of the gas accreting onto, or flowing out of, young galaxies. Lyman-α studies may therefore ultimately hold the key to understanding the initial stages of galaxy formation and growth.

Chapter Eleven

The 21-cm Line

As powerful as it is, the Lyman-α transition has several major disadvantages for studying the high-z Universe:

- Most importantly, the Gunn-Peterson optical depth is enormous. Even a very small neutral fraction, of order $\sim 10^{-3}$, suffices to render the IGM opaque in this line. Thus, we are not able to use it to study the early, or even middle phases of reionization except in special circumstances.
- Because Lyman-α absorption is in the UV band, observing it requires bright UV sources which are very rare at high redshifts, and limits the related studies to only rare redshift skewers (lines of sight).
- The high excitation energy of the Lyman-α transition prevents us from using it to study the cold pre-reionization IGM, because the temperatures are much too low there to collisionally excite the line. Moroever, the large optical depth for absorption prevents us from measuring the IGM temperature through the line width.

The first of these can be remedied by using a resonant transition of a rarer element, such as metals, but of course such elements are rare, and their distribution introduces extra uncertainty into the interpretation. We can address all these problems by searching for a weaker, lower-energy line of atomic hydrogen: the best candidate is the spin-flip, or hyperfine line. This transition was predicted by Hendrik van de Hulst in 1944 (following a suggestion by Oort) and first observed from the sky by Harold Ewen and Ed Purcell through an office window at the Harvard Physics department in 1951. It is driven by the interaction of the magnetic moments of the proton and electron; when these moments are aligned, the atom has a slightly higher energy than when they are anti-aligned. An atom in the upper state will then eventually undergo a "spin-flip" transition, emitting a photon with a wavelength of 21 cm. As we shall see, this transition is extremely weak, so the effective IGM optical depth is only of order 1%: this makes the entire neutral IGM accessible during the "cosmic dawn." Moreover, the transition energy is so low that it provides a sensitive thermometer of the low-temperature IGM, and – as a low-frequency radio transition – it can be seen across the entirety of the IGM against the CMB.

Figure 11.1 illustrates the power of the spin-flip transition with an analogy to the well-known structure of "Swiss cheese". Each slice of cheese has a different structure, depending on where the air bubbles happen to lie within it. In the case of the spin-flip transition, by observing different wavelengths of 21 cm$\times(1+z)$,

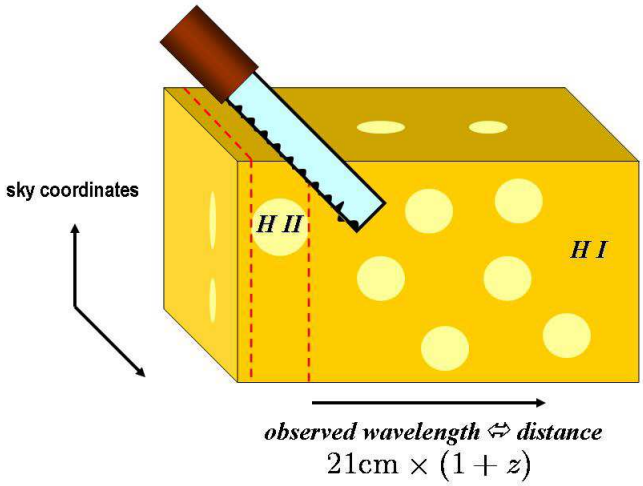

Figure 11.1 21-cm imaging of ionized bubbles during the epoch of reionization is analogous to slicing Swiss cheese. The technique of slicing at intervals separated by the typical dimension of a bubble is optimal for revealing different patterns in each slice.

one is slicing the Universe at different redshifts z. Moreover, the redshifted 21-cm emission should display angular structure as well as frequency structure due to inhomogeneities in the gas density, the hydrogen ionized fraction, and the fraction of excited atoms – the analog of the air bubbles in Swiss cheese. A full map of the distribution of H I as a function of redshift would provide a three-dimensional image of the Swiss-cheese structure of the IGM during reionization. This mapping "tomography" provides the only way to map the distribution of $> 90\%$ of the matter in the Universe during the Dark Ages and cosmic dawn.

Figure 11.2 shows a more concrete overview of the expected spin-flip signal. This has two interesting aspects: the sky-averaged, or monopole, brightness, which records the average properties of the H I as a function of observed wavelength (or equivalently cosmic time). This is shown in the bottom panel in brightness temperature units relative to the CMB (see below for a detailed discussion). Several different phases are labeled; we will discuss each in turn in this chapter. The top panel shows the fluctuations inherent in this signal, which arise from the discrete, clustered luminous sources. The spin-flip background measures the ultraviolet and X-ray radiation fields over a broad swath of cosmic history, complementing the direct probes of individual galaxies that we have already described.

This chapter will describe how we use the 21-cm line to study the high-z Universe. Following convention in the literature, we will often refer to the signal as the "21-cm radiation," although in reality the *observed* wavelengths are larger by a factor of $(1 + z)$.

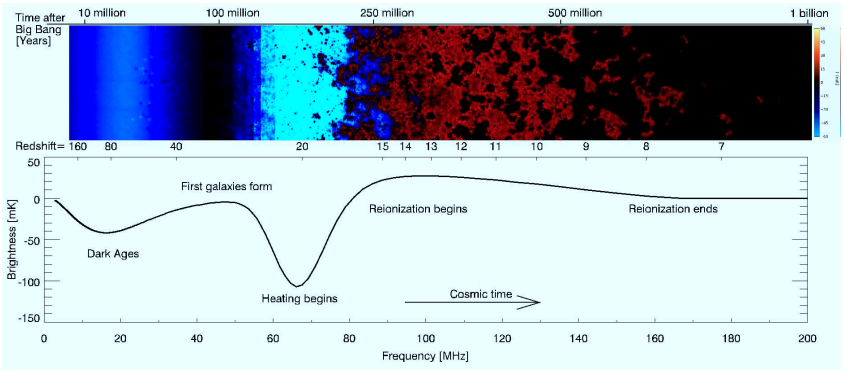

Figure 11.2 Overview of the expected 21 cm signal. *Top panel:* Time evolution of fluctuations in the 21 cm brightness from just before the first stars form through to the end of reionization. This evolution is pieced together from instantaneous redshift slices through a (100 Mpc)^3 numerical simulation volume. Coloration indicates the strength of the 21 cm brightness as it transitions from absorption (blue) to emission (red) and finally disappears (black) due to ionization. *Bottom panel:* Expected evolution of the sky-averaged 21cm brightness from the "Dark Ages" at $z = 150$ to the end of reionization sometime before $z = 6$. The frequency structure is driven by the interplay of gas heating, the coupling of gas and 21 cm temperatures, and the ionization of the gas. There is considerable uncertainty in the exact form of this signal arising from the poorly understood properties of the first galaxies. Figure credit: J. Pritchard & A. Loeb, Nature **468**, 772 (2010).

11.1 RADIATIVE TRANSFER OF THE 21-CM LINE

The radiative transfer equation for the specific intensity I_ν of a spectral line reads

$$\frac{dI_\nu}{d\ell} = \frac{\phi(\nu)h\nu}{4\pi}\left[n_2 A_{10} - (n_1 B_{01} - n_2 B_{10})\, I_\nu\right], \tag{11.1}$$

where $d\ell$ is a path length 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 and centered around the line frequency, subscripts 1 and 2 denote the lower and upper atomic levels, n denotes the number density of atoms at the different levels, and A and B are the Einstein coefficients for the transition between these levels. In our case, the line frequency $\nu_{21} = 1420.4057$ MHz corresponds to a

wavelength of $\lambda_{21} = 21.1061$ cm. We can then make use of the standard relations in atomic physics: $B_{10} = (g_1/g_2)B_{01}$ and $B_{10} = A_{10}(c^2/2h\nu^3)$, where g is the spin degeneracy factor of each state. For the 21-cm transition, $A_{10} = 2.85 \times 10^{-15}\mathrm{s}^{-1}$ and $g_2/g_1 = 3$.

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

$$\left(\frac{n_2}{n_1}\right) = \left(\frac{g_2}{g_1}\right) \exp\left\{-\frac{E_{10}}{k_B T_S}\right\}, \tag{11.2}$$

where $T_* \equiv E_{10}/k_B = 68$ mK is equivalent to the transition energy. In the regime of interest, E/k_B is much smaller than the CMB temperature T_γ as well as the spin temperature T_S, and so all related exponentials can be expanded to leading order.

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 $T_b(\nu) \approx I_\nu\, c^2/2k_B\nu^2$.

In the Rayleigh-Jeans limit, the equation of radiative transfer along a line of sight through a cloud of uniform excitation temperature T_S becomes

$$T_b'(\nu) = T_s(1 - e^{-\tau_\nu}) + T_R'(\nu)e^{-\tau_\nu} \tag{11.3}$$

where the *optical depth* $\tau_\nu \equiv \int ds\, \alpha_\nu$ is the integral of the absorption coefficient (α_ν) along the ray through the cloud, T_R' is the brightness of the background radiation field incident on the cloud along the ray, and s is the proper distance. 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 $\nu = \nu_0/(1+z)$. Henceforth we will work in terms of these observed quantities.

The absorption coefficient is determined from the Einstein coefficients via

$$\alpha = \phi(\nu)\frac{h\nu}{4\pi}[n_1 B_{01} - n_2 B_{10}). \tag{11.4}$$

Because all astrophysical applications have $T_S \gg T_*$, approximately three of four atoms find themselves in the excited state ($n_2 \approx n_1/3$). As a result, the stimulated emission correction is significant (and the net absorption depends on T_S).

In an expanding Universe with a local hydrogen number density $n_{\rm H}$ and with a velocity gradient along the line of sight of $dv_\parallel/dr_\parallel$, the 21-cm optical depth can be derived similarly to equation(4.13).[i] Writing $\phi(\nu) \sim 1/(\Delta\nu)$ we get

$$\tau_{10} = \frac{3}{32\pi}\,\frac{hc^3 A_{10}}{k_B T_S \nu_{10}^2}\,\frac{x_{\rm HI} n_{\rm H}}{(1+z)\,(dv_\parallel/dr_\parallel)} \tag{11.5}$$

$$\approx 0.0092\,(1+\delta)\,(1+z)^{3/2}\,\frac{x_{\rm HI}}{T_S}\,\left[\frac{H(z)/(1+z)}{dv_\parallel/dr_\parallel}\right], \tag{11.6}$$

In the second part we express T_S in kelvin and have scaled to the mean IGM density at z and to the average velocity gradient (the Hubble flow). In the latter case,

[i] Interestingly, the 21 cm case was actually computed by George Field in 1959, several years *before* the Gunn-Peterson calculation.

$\Delta I_\nu \propto \Delta\ell\phi(\nu)\nu = |cdt/dz|(\nu dz/d\nu) = c/H$, providing the analog of the Gunn-Peterson optical depth.

In practice, the background radiation source is usually the CMB, so $T_R' = T_\gamma(z)$, and we are observing the contrast between high-redshift hydrogen clouds and the CMB. Because the optical depth is so small, we can expand the exponentials in equation (11.3), which yields

$$T_b(\nu) \approx \frac{T_S - T_\gamma(z)}{1+z}\,\tau_{\nu_0} \tag{11.7}$$

$$\approx \quad 9\,x_{\rm HI}(1+\delta)\,(1+z)^{1/2}\left[1-\frac{T_\gamma(z)}{T_S}\right]\left[\frac{H(z)/(1+z)}{dv_\parallel/dr_\parallel}\right]\ {\rm mK}. \tag{11.8}$$

Here $T_b < 0$ if $T_S < T_\gamma$, yielding an absorption signal, or emission otherwise; both regimes are important for the high-z Universe. Note that δT_b saturates if $T_S \gg T_\gamma$, but the absorption can become arbitrarily large if $T_S \ll T_\gamma$. The observability of the 21 cm transition therefore hinges on the spin temperature; we will next describe the mechanisms that drive T_S either above or below $T_\gamma(z)$.

11.2 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. In the presence of the CMB alone, the spin states reach thermal equilibrium with $T_S = T_\gamma$ on a time-scale of $\sim T_*/(T_\gamma A_{10}) = 3\times10^5(1+z)^{-1}$ yr, where $A_{10} = 2.87\times10^{-15}$ s^{-1} is the spontaneous decay rate of the hyperfine transition. This time scale is much shorter than the age of the Universe at all redshifts after cosmological recombination. However, the other two processes break this coupling. We let C_{10} and P_{10} be the de-excitation rates (per atom) from collisions and UV scattering, respectively. We also let C_{01} and P_{01} be the corresponding excitation rates. The spin temperature is then determined in equilibrium by[ii]

$$n_1\,(C_{10} + P_{10} + A_{10} + B_{10}I_{\rm CMB}) = n_0\,(C_{01} + P_{01} + B_{01}I_{\rm CMB})\,, \tag{11.9}$$

where $I_{\rm CMB}$ is the specific intensity of CMB photons. With the Rayleigh-Jeans approximation, equation (11.9) can be rewritten as

$$T_S^{-1} = \frac{T_\gamma^{-1} + x_c T_K^{-1} + x_\alpha T_c^{-1}}{1 + x_c + x_\alpha}, \tag{11.10}$$

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 the principle of detailed balance through the relation

$$\frac{C_{01}}{C_{10}} = \frac{g_2}{g_1}e^{-T_\star/T_K} \approx 3\left(1-\frac{T_\star}{T_K}\right). \tag{11.11}$$

[ii]Note that the relevant timescales are all much shorter than the expansion time, so equilibrium is an excellent approximation.

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

$$\frac{P_{01}}{P_{10}} \equiv 3\left(1 - \frac{T_\star}{T_c}\right). \tag{11.12}$$

We will next calculate x_c, x_α, and T_c. In the limit in which $T_c \to T_K$ (a reasonable approximation in most situations of interest), equation (11.10) may be written as

$$1 - \frac{T_\gamma}{T_S} = \frac{x_c + x_\alpha}{1 + x_c + x_\alpha}\left(1 - \frac{T_\gamma}{T_K}\right). \tag{11.13}$$

11.2.1 Collisional Coupling

We will first consider collisional excitation and de-excitation of the hyperfine levels, which dominate in dense gas. The coupling coefficient for species i is

$$x_c^i \equiv \frac{C_{10}^i}{A_{10}}\frac{T_\star}{T_\gamma} = \frac{n_i\,\kappa_{10}^i}{A_{10}}\frac{T_\star}{T_\gamma}, \tag{11.14}$$

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, which in principle includes collisions with (1) other hydrogen atoms, (2) free electrons, (3) protons, and (4) other species (helium and deuterium); the last turn out to be unimportant.

These rate coefficients are ultimately determined by the quantum mechanical cross sections of the relevant processes; we will not list them in detail but merely present the results in Figure 11.3. Although the atomic cross-section is small, in the unperturbed IGM collisions between neutral hydrogen atoms nearly always dominate these rates because the ionized fraction is small. Free electrons can be important in partially ionized gas, especially at relatively high temperatures; collisions with protons are only important at the lowest temperatures.

Crucially, the collisional coupling is quite weak in a nearly neutral, cold medium. Thus, the overall density must be large in order for this process to effectively fix T_S. A convenient estimate of their importance is the critical overdensity, $\delta_{\rm coll}$, at which $x_c = 1$:

$$1 + \delta_{\rm coll} = 1.06\left[\frac{\kappa_{10}(88\ \mathrm{K})}{\kappa_{10}(T_K)}\right]\left(\frac{0.023}{\Omega_b h^2}\right)\left(\frac{70}{1+z}\right)^2, \tag{11.15}$$

where we have inserted the expected temperature at $1 + z = 70$. Thus for redshifts $z < 70$, $T_S \to T_\gamma$; by $z \sim 30$ the IGM essentially becomes invisible. It is worth emphasizing that κ_{10} is extremely sensitive to T_K in this regime. If the universe is somehow heated above the fiducial value, the threshold density can remain modest: $\delta_{\rm coll} \approx 1$ at $z = 40$ if $T_K = 300\ \mathrm{K}$.

11.2.2 The Wouthuysen-Field Effect

We therefore require a different mechanism to break the coupling to the CMB during the era of the first galaxies. This is known as the Wouthuysen-Field mechanism

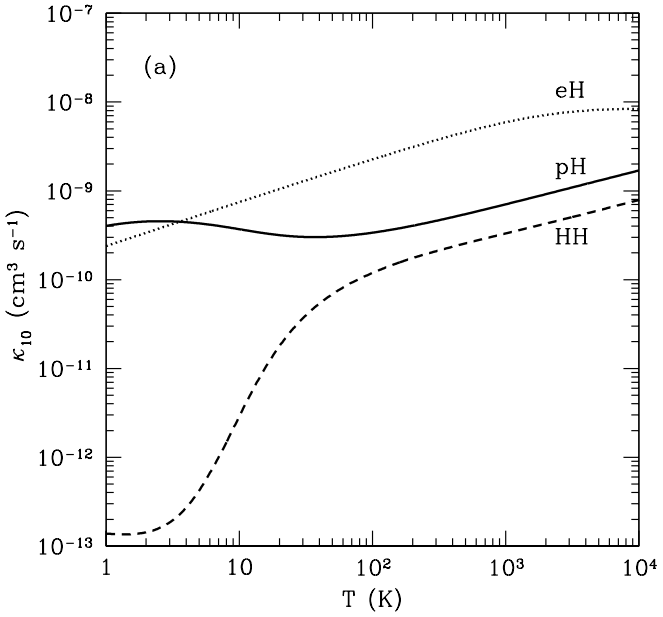

Figure 11.3 De-excitation rate coefficients for H-H collisions (dashed line), H-e$^-$ collisions (dotted line), and H-p collisions (solid line). Note that the net rates are also proportional to the densities of the individual species, so H-H collisions still dominate in a weakly-ionized medium. Figure credit: S. R. Furlanetto & M. R. Furlanetto, *Mon. Not. R. Astron. Soc.* **379**, 130 (2007).

(named after the Dutch physicist Siegfried Wouthuysen and Harvard astrophysicist George Field who explored it first[40], and it is illustrated in Figure 11.4, where we have drawn the hyperfine sub-levels of the $1S$ and $2P$ states of HI. Suppose a hydrogen atom in the hyperfine singlet state absorbs a Lyman-α photon. The electric dipole selection rules allow $\Delta F = 0, 1$ except that $F = 0 \rightarrow 0$ 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.[iii] Thus, atoms can change hyperfine states through the absorption and spontaneous re-emission of a Lyman-α photon (or indeed any Lyman-series photon).

The Wouthuysen-Field coupling must depend on the total rate (per atom) at which Lyman-α photons are scattered within the gas,

$$P_\alpha = 4\pi\chi_\alpha \int d\nu\, J_\nu(\nu)\phi_\alpha(\nu), \tag{11.16}$$

where $\sigma_\nu \equiv \chi_\alpha\phi_\alpha(\nu)$ is the local absorption cross section, $\chi_\alpha \equiv (\pi\, e^2/m_e\, c)f_\alpha$,

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

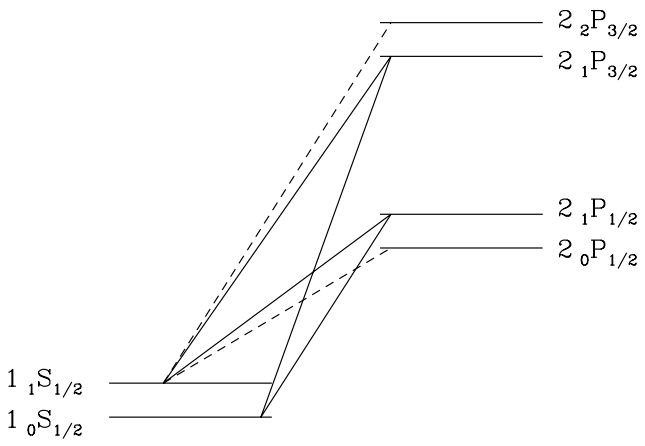

Figure 11.4 Level diagram illustrating the Wouthuysen-Field effect. We show the hyperfine splittings of the $1S$ and $2P$ levels. The solid lines label transitions that mix the ground state hyperfine levels, while the dashed lines label complementary transitions that do not participate in mixing. Figure credit: J. R. Pritchard & S. R. Furlanetto *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006).

$f_\alpha = 0.4162$ is the oscillator strength of the Lyman-α transition, $\phi_\alpha(\nu)$ is the Lyman-α absorption profile, and J_ν is the angle-averaged specific intensity of the background radiation field.[iv] The line typically has a Voigt profile ϕ_V as described in §10.1.1.

What about transitions to higher Lyman-n levels? Suppose that a photon redshifts into the Lyman-n resonance. After absorption, it can either scatter (through a decay directly to the ground state) or cascade through a series of intermediate levels and produce different photons. The direct decay probabilities are ~ 0.8, so a Lyman-n photon will typically scatter $N_{\rm scatt} \approx (1 - P_{nP\rightarrow 1S})^{-1} \sim 5$ times before instead initiating a decay cascade. As a result, coupling from the direct scattering of Lyman-n photons is suppressed compared to Lyman-α by a large factor.

However, Lyman-n photons can still be important because of their cascade products. Consider the decay chains shown in Figure 11.5. For Lyβ, the only permitted decays are to the ground state (regenerating a Lyβ photon and starting the process again) or to the $2S$ level. The Hα photon produced in the $3P \rightarrow 2S$ transition

[iv] By convention, we use the specific intensity in units of photons cm^{-2} Hz^{-1} s^{-1} sr^{-1} here, which is conserved during the expansion of the Universe (whereas energy redshifts away).

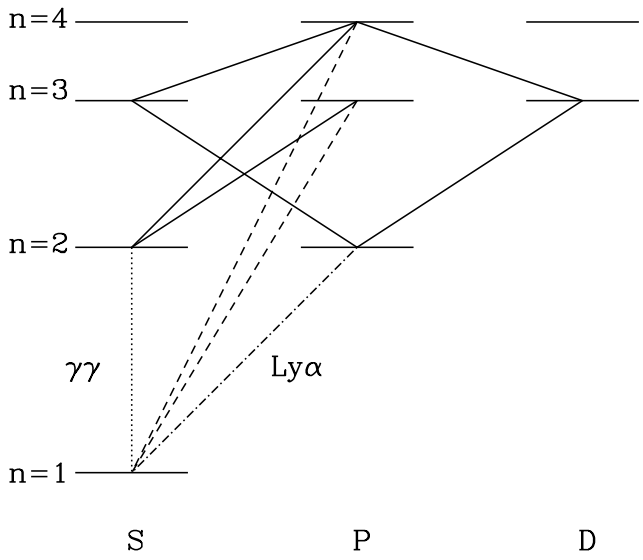

Figure 11.5 Decay chains for Lyβ and Lyγ. We show Lyman-n transitions by dashed curves, Lyman-α by the dot-dashed curve, cascades by solid curves, and the forbidden $2S \rightarrow 1S$ transition by the dotted curve. Figure credit: J. R. Pritchard & S. R. Furlanetto, *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006).

(and indeed any photon produced in a decay to an excited state) escapes to infinity. Thus the atom will eventually find itself in the $2S$ state, which decays to the ground state via a forbidden two photon process with $A_{2S\rightarrow 1S} = 8.2\ \mathrm{s}^{-1}$. These photons too will escape to infinity. Thus coupling from Lyβ photons can be completely neglected.

But now consider excitation by Lyγ, also shown in Figure 11.5. This can cascade (through $3S$ or $3D$) to the $2P$ level, in which case the original Lyman-n photon is "recycled" into a Lyman-α photon, which then scatters many times through the IGM. Thus, the key quantity for determining the coupling induced by Lyman-n photons is the fraction $f_{\rm rec}(n)$ of cascades that terminate in Lyman-α photons. We have seen that $f_{\rm rec}(n = 3)$ vanishes, but the higher states all have $f_{\rm rec} \sim 1/3$.

Focusing on the Lyman-α photons themselves, we must relate the total scattering rate P_α to the indirect de-excitation rate P_{10}. We will make the simplifying assumption that the specific intensity J_ν is constant across the Lyman-α transition. 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 $i \rightarrow j$ transition as J_{ij}. Then

$$P_{01} \propto B_{\rm ad} J_{\rm ad} \frac{A_{\rm db}}{A_{\rm da} + A_{\rm db}} + B_{\rm ae} J_{\rm ae} \frac{A_{\rm eb}}{A_{\rm ea} + A_{\rm eb}}. \tag{11.17}$$

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 Lyman-α 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 letter labels 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

$$x_\alpha = \frac{4P_\alpha}{27A_{10}} \frac{T_\star}{T_\gamma} = S_\alpha \frac{J_\alpha}{J_\nu^c}, \tag{11.18}$$

where in the second equality we evaluate J_ν at line center and set $J_\nu^c \equiv 1.165 \times 10^{-10}[(1+z)/20]$ cm^{-2} s^{-1} Hz^{-1} sr^{-1}. We include here a correction factor S_α that accounts for variations in the intensity near the line center (see below). This coupling threshold for $x_\alpha = S_\alpha$ can also be written in terms of the number of Lyman-α photons per hydrogen atom, which we denote $\tilde{J}_\nu^c = 0.0767\,[(1+z)/20]^{-2}$. This threshold is relatively easy to achieve in practice.

Two challenges remain: calculating T_c and the correction factor S_α. The former is the effective temperature of the UV radiation field, defined in equation (11.12), which is determined by the shape of the photon spectrum at the Lyman-α resonance. That the effective temperature of the radiation field *must* matter is easy to see: the energy deficit between the different hyperfine splittings of the Lyman-α transition implies that the mixing process is sensitive to the gradient of the background spectrum near the Lyman-α resonance. More precisely, the procedure described near equation (11.17) yields

$$\frac{P_{01}}{P_{10}} = \frac{g_1}{g_0} \frac{n_{\rm ad} + n_{\rm ae}}{n_{\rm bd} + n_{\rm be}} \approx 3 \left(1 + \nu_0 \frac{d\ln n_\nu}{d\nu}\right), \tag{11.19}$$

where $n_\nu = c^2\, J_\nu / 2\nu^2$ is the photon occupation number. Thus, by comparison to equation (11.12) we find

$$\frac{h}{k_B T_c} = -\frac{d\ln n_\nu}{d\nu}. \tag{11.20}$$

A simple argument shows that $T_c \approx T_K$: so long as the medium is extremely optically thick, the enormous number of Lyman-α scatterings must bring the Lyman-α profile to a blackbody of temperature T_K near the line center. This condition is easily fulfilled in the high-redshift IGM, where $\tau_{\rm GP} \gg 1$. In detail, atomic recoils during scattering tilt the spectrum to the red and are primarily responsible for establishing this equilibrium.

The scattering process is actually much more complicated than naively expected because scattering itself modifies the shape of J_ν. (At low temperatures, where the line broadening is small, the hyperfine sub-levels of the Lyman-α transition must also be taken into account.) Intuitively, a flat input spectrum develops an absorption feature because of the increased scattering rate near the Lyman-α resonance. Photons continually lose energy by redshifting, but they also lose energy through recoil whenever they scatter. If the fractional frequency drift is denoted by $\mathcal{A}$, continuity requires $n_\nu \mathcal{A}$=constant; when $\mathcal{A}$ increases near resonance, the number density must fall. On average, the energy loss (or gain) per scattering is

$$\frac{\Delta E_{\rm recoil}}{E} = \frac{h\nu}{m_p c^2}\left(1 - \frac{T_K}{T_c}\right), \tag{11.21}$$

where the first factor comes from recoil off an isolated atom and the second factor corrects for the distribution of initial photon energies; the energy loss vanishes when $T_c = T_K$, and when $T_c < T_K$, the gas is heated by the scattering process.

To compute the suppression factor in the intensity we must calculate the photon spectrum near Lyman-α. We begin with the radiative transfer equation in an expanding universe (written in comoving coordinates, and again using units of cm^{-2} s^{-1} Hz^{-1} sr^{-1} for J_ν):

$$\frac{1}{c n_H \chi_\alpha}\frac{\partial J_\nu}{\partial t} = -\phi_\alpha(\nu)\, J_\nu + H\nu_\alpha \frac{\partial J_\nu}{\partial \nu} + \int d\nu'\, R(\nu,\nu')\, J_{\nu'} + C(t)\psi(\nu). \tag{11.22}$$

Here the first term on the right-hand side describes absorption, the second involves the Hubble flow, and the third accounts for re-emission following absorption. $R(\nu,\nu')$ is the "redistribution function" that describes the frequency of an emitted photon, which depends on the relative momenta of the absorbed and emitted photons as well as the absorbing atom. The last term describes injection of new photons: C is the rate at which they are produced and $\psi(\nu)$ is their frequency distribution.

The redistribution function R is the complicated aspect of the problem, but it can be simplified if the frequency change per scattering (typically of order $\Delta\nu_D$) is "small." In that case, we can expand $J_{\nu'}$ to second order in $(\nu - \nu')$ and rewrite equation (11.22) as a diffusion problem in frequency. The steady-state version of equation (11.22) becomes, in this so-called *Fokker-Planck* approximation,

$$\frac{d}{dx}\left(-\mathcal{A}\, J + \mathcal{D}\frac{dJ}{dx}\right) + C\psi(x) = 0, \tag{11.23}$$

where $x \equiv (\nu - \nu_\alpha)/\Delta\nu_D$, and $\mathcal{D}$ is the diffusivity. In general, the Fokker-Planck approximation is valid when (i) the frequency change per scattering ($\sim \Delta\nu_D$) is smaller than the width of any spectral features, and either (iia) the photons are outside the line core where $d\phi_\alpha/dx$ is small, or (iib) the atoms are in equilibrium with $T_c \approx T_K$.

Solving for the background spectrum thus reduces to specifying $\mathcal{A}$ and $\mathcal{D}$. The first involves the Hubble flow, which causes a drift $\mathcal{A}_H = -\tau_{\rm GP}^{-1}$ (without any associated diffusion). The remaining terms come from R and incorporate all the physical processes relevant to energy exchange in scattering. The drift from recoil

is

$$\mathcal{D}_{\rm scatt} = \phi_\alpha(x)/2, \tag{11.24}$$

$$\mathcal{A}_{\rm scatt} = -(\eta - x_0^{-1})\phi_\alpha(x), \tag{11.25}$$

where $x_0 \equiv \nu_\alpha/\Delta\nu_D$ and $\eta \equiv (h\nu_\alpha^2)/(m_p c^2 \Delta\nu_D)$. The latter is the recoil parameter measuring the average loss per scattering in units of the Doppler width.

Finally, to solve equation (11.23) we must specify the boundary conditions, which essentially correspond to the input photon spectrum (ignoring scattering) and the source function. Because the frequency range of interest is so narrow, two cases suffice: a flat input spectrum (which approximately describes photons that redshift through the Lyman-α resonance, regardless of the initial source spectrum) and a step function, where photons are "injected" at line center (through cascades or recombinations) and redshift away. In either case, the first integral over x is trivial. At high temperatures where spin flips are unimportant to the overall energy exchange, we can write

$$\phi\frac{dJ}{dx} + 2\{[\eta - (x + x_0)^{-1}]\phi + \tau_{\rm GP}^{-1}\}J = 2K/\tau_{\rm GP}. \tag{11.26}$$

The integration constant K equals J_∞, the flux far from resonance, for photons that redshift into the line and for injected photons at $x < 0$; it is zero for injected photons at $x > 0$. (In practice, spin flips affect both the scattering rate and color temperature at small temperatures, modifying both the drift and diffusion terms.)

The formal analytic solution, when $K \neq 0$, is most compactly written in terms of $\delta_J \equiv (J_\infty - J)/J_\infty$:

$$\delta_J(x) = 2\eta \int_0^\infty dy \exp\left[-2\eta' y - 2\gamma' \int_{x-y}^{x} \frac{dx'}{\phi_\alpha(x')}\right]. \tag{11.27}$$

(An analogous form also exists for photons injected at line center.) The full problem, including the intrinsic Voigt profile of the Lyman-α line, must be solved numerically, but including only the Lorentzian wings from natural broadening allows a simpler solution. Fortunately, this assumption is quite accurate in the most interesting regime of $T_K < 1000$ K.

The crucial aspect of equation (11.27) is that (as expected from the qualitative argument above) an absorption feature appears near the line center; its strength is roughly proportional to η, our recoil parameter. The feature is more significant when T_K is small (or the average effect of recoil is large). Figure 11.6 shows some example spectra (both for a continuous background and for photons injected at line center).

For now, the most important result is the suppression of the radiation spectrum at line center compared to the assumed initial condition. This decreases the total scattering rate of Lyman-α photons (and hence the Wouthuysen-Field coupling) below what one naively expects. The suppression factor is

$$S_\alpha = \int_{-\infty}^{\infty} dx\, \phi_\alpha(x)\, J(x) \approx [1 - \delta_J(0)] \leq 1, \tag{11.28}$$

where the second equality follows from the narrowness of the line profile. Again, the Lorentzian wing approximation turns out to be an excellent one; when spin

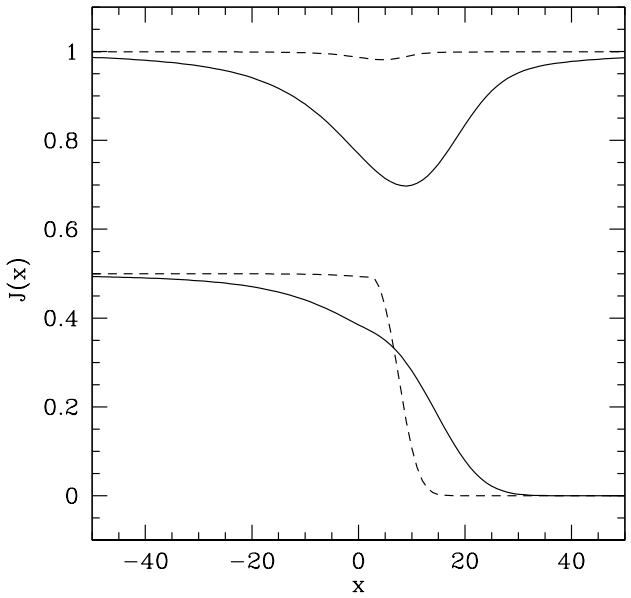

Figure 11.6 Background radiation field near the Lyman-α resonance at $z = 10$; $x \equiv (\nu - \nu_\alpha)/\Delta\nu_D$ is the normalized deviation from line center. The upper and lower sets are for continuous photons and photons injected at line center, respectively. (The former are normalized to J_∞; the latter have arbitrary normalization.) The solid and dashed curves take $T_K = 10$ and 1000 K, respectively. Figure credit: S. R. Furlanetto & J. R. Pritchard, *Mon. Not. R. Astron. Soc.* **372**, 1093 (2006).

exchange can be neglected, the suppression is

$$S_\alpha \sim \exp\left[-0.803 \left(\frac{T_K}{1\mathrm{K}}\right)^{-2/3} \left(\frac{\tau_{\mathrm{GP}}}{10^6}\right)^{1/3}\right]. \tag{11.29}$$

Note that this form applies to both photons injected at line center as well as those that redshift in from infinity. As we can see in Figure 11.6, the suppression is most significant in cool gas.

11.3 THE BRIGHTNESS TEMPERATURE OF THE SPIN-FLIP BACKGROUND

With the basic atomic physics of the 21-cm line in place, we now turn to estimating the astrophysical inputs that determine its properties. Of course, these inputs are at the moment unknown, so we will at first keep the discussion general and then later focus on some particular simple models.

11.3.1 Feedback: The Lyman-α Background

After $z \sim 30$, when collisional coupling becomes unimportant, the spin temperature is determined by the scattering of Lyman-α photons. In practice, the relevant photons do not start at the Lyman-α wavelength, because those redshift out of resonance very soon after they are created and do not contribute to the coupling except very near their sources. Instead, these photons begin in the ultraviolet and redshift into a Lyman-series line, possibly cascading down to a Lyman-α photon.

To compute J_α, we therefore begin with the comoving ultraviolet emissivity at a frequency ν, $\epsilon(\nu, z)$. Here we will consider the simple limit in which this emissivity is nearly uniform.

In fact, we have already discussed this background in some detail, for these photons, which range in energy from 10.2–13.6 eV, are (nearly) the same as those which contribute to the Lyman-Werner background that dissociates H_2 molecules in the early Universe. The difference is that we are concerned not with the photons between the Lyman resonances but those photons that do redshift into those resonances (and then cascade into Lyman-α, in the case of higher-n transitions). Given the ultraviolet emissivity, the seeked-after background is

$$\begin{aligned} J_\alpha(z) &= \sum_{n=2}^{n_{\rm max}} J_\alpha^{(n)}(z) \\ &= \sum_{n=2}^{n_{\rm max}} f_{\rm rec}(n) \int_z^{z_{\rm max}(n)} dz' \, \frac{(1+z)^2}{4\pi} \, \frac{c}{H(z')} \, \epsilon(\nu_n', z'), \end{aligned} \quad (11.30)$$

where ν_n' is the frequency at redshift z' that redshifts into the Lyman-n resonance at redshift z, and $z_{\rm max}(n)$ is the largest redshift from which a photon can redshift into the Lyman-n resonance. The sum must be truncated at some large $n_{\rm max}$ that is determined by the typical size of ionized regions around the sources, but the result is not sensitive to its precise value.

Just as with the Lyman-Werner background, the Lyman-α intensity is quite uniform: in fact the effective "horizon" within which a given source is visible is even larger than in that other case, because the gap between Lyman-α and Lyman-β corresponds to ~ 250 comoving Mpc. However, unlike for the Lyman-Werner background, Wouthuysen-Field coupling is rather sensitive to the precise intensity of the background, so the fluctuations are still very important.

These, in turn, depend on the sources of the emissivity. The most obvious sources are star-forming galaxies. If the star formation rate traces the rate at which matter collapses into galaxies, the comoving emissivity at frequency ν is

$$\epsilon(\nu, z) = f_\star \, \bar{n}_b^c \, \epsilon_b(\nu) \, \frac{df_{\rm coll}}{dt}, \quad (11.31)$$

where $\epsilon_b(\nu)$ is the number of photons produced in the frequency interval $\nu \pm d\nu/2$ per baryon incorporated into stars. Although real spectra are rather complicated, a useful quantity is the total number N_α of photons per baryon in the interval 10.2–13.6 eV. For low-metallicity Pop II stars and very massive Pop III stars, this is $N_\alpha = 9690$ and $N_\alpha = 4800$, respectively.

Of course, processes other than star formation can also create a Lyman-α background. These include especially UV photons from quasars (which can be modeled in the same way at stars, though ϵ changes) and collisional excitation by higher energy X-rays. For the latter, a fraction $f_c \sim f_i \sim x_{\rm HI}/3$ of the energy is typically lost to excitations, and ≈ 0.8 of that energy ends up in Lyman-α photons. We call this total rate f_α, and it is easy to see that it can be quite significant. The critical intensity for the Wouthuysen-Field effect corresponds to ~ 1 photon per ~ 10 hydrogen atoms, or ~ 1 eV per atom. Thus, in any scenario that appeals to X-rays for significant ionization, we would expect strong coupling once the IGM becomes $\sim 10\%$ ionized.

In any case, computing the fluctuations in the intensity is a more difficult task. The simplest approach is to use a modified version of the halo model (introduced in §3.7.1) applied to the radiation background instead of the density field. Here we construct the background by imagining that each galaxy is surrounded by a radiation field with a specified shape, $J_h(r|\Psi)$, where Ψ labels all the parameters that may determine an individual galaxy's luminosity (principally, we will presume below, the host halo's mass). The total radiation background $J(\mathbf{x})$ is then the sum of that from each halo, just as the density field is the sum of the density profiles of each dark matter clump in the usual halo model. We can therefore use the usual machinery of the halo model to describe the radiation background. Then

$$P_J(k) = P_J^{1h}(k) + P_J^{2h}(k), \tag{11.32}$$

where P_J^{1h} describes correlations from a single galaxy's radiation and P_J^{2h} those between different galaxies.

The key input is therefore to determine the intensity profile of each galaxy. If we consider only the radiation between Lyman-α and Lyman-β, the profile of photons that redshift into the Lyman-α resonance will follow the usual $1/r^2$ law with just two modifications: (1) the profile will be truncated where those photons with the largest initial energies (just below Lyman-β) redshift into the Lyman-α resonance, and (2) the relevant emitted frequency (chosen so that it redshifts into Lyman-α resonance at r) varies with radius r. However, we must also add those photons which redshift into a higher Lyman-n series resonance and then cascade to Lyman-α. Thus, the total profile is

$$J_{h,\alpha}(r) = \sum_{n=1}^{\infty} f_{\rm rec}(n) \frac{L(\nu_n'|\Psi)/h\nu_n'}{(4\pi r)^2}, \tag{11.33}$$

where $L(\nu|\Psi)$ is the luminosity per unit frequency from the source (with parameters Ψ) and ν_n' is the frequency that redshifts into the Lyman-n resonance at r and each term in the sum is only included when r is smaller than the effective horizon for these photons. The left panel in Figure 11.7 shows this profile for a massive galaxy at $z = 20$. Note that it is slightly steeper than the $1/r^2$ expectation at moderate distances from the source, owing to the cascade effects.

In practice, the light travel time over the Lyman-α horizon is > 100 Myr, a substantial fraction of both the age of Universe and the lifetime of a typical source (either stars or quasars). As a result, the source luminosity likely changes significantly during the time period of interest, and some estimate of source evolution is

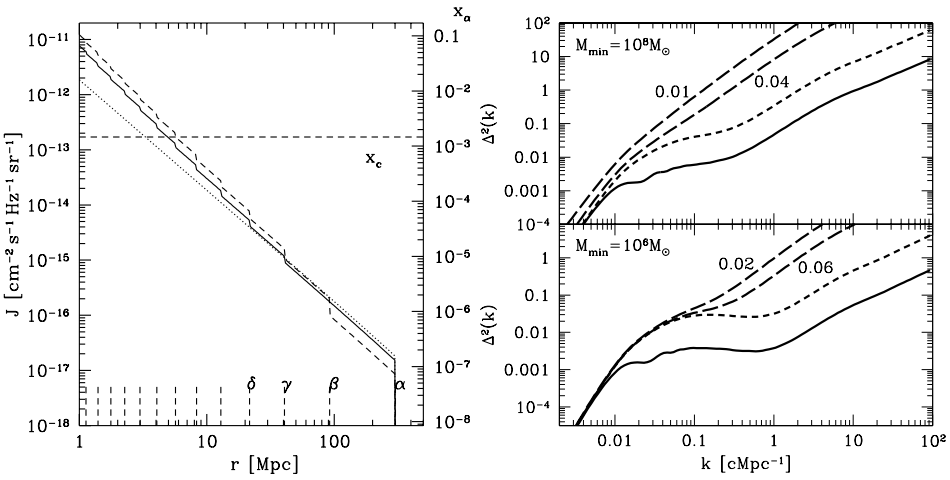

Figure 11.7 *Left:* Lyman-α flux profile of a massive galaxy at $z = 20$. The solid line shows the full calculation the dashed line assumes $f_{\rm rec} = 1$ for all n, and the dotted line ignores all cascades from higher energies (all are normalized to have the same total luminosity). The vertical dashed lines along the horizontal axis mark the horizons of the respective Lyman transitions. The right axis converts the local flux to the Wouthuysen-Field coupling coefficient assuming $S_\alpha = 1$; for context, the dashed horizontal line shows the collisional coupling coefficient at $z = 20$ assuming no IGM heating. *Right:* Dimensionless power spectrum of the Lyman-α background for several simple star formation scenarios. The solid curves are chosen to be near the peak of the Lyman-α fluctuations. They assume $M_{\rm min} = 10^6\ M_\odot$ at $z = 30$ (bottom curve) and $M_{\rm min} = 10^8\ M_\odot$ at $z = 20.5$ (top curve); both take $f_\star = 0.1$ to normalize the background. The dashed curve takes $M_{\rm min} = 10^6\ M_\odot$ at $z = 30$ but assumes that each halo can only form stars for 3 Myr, so that $f_{\rm duty} = 0.02$. Figure credit: J. R. Pritchard & S. R. Furlanetto *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006); L. Holzbauer & S. R. Furlanetto, *Mon. Not. R. Astron. Soc.*, submitted (2011).

necessary. In the simplest models, one can take $L \propto df_{\rm coll}/dt$ in order to reflect the overall evolution of gas inside galaxies.

The right panel of Figure 11.7 shows some example power spectra for the intensity of this background. We consider two cases: one in which the luminous sources are extremely small (existing in all halos above $10^6\ M_\odot$, solid curves) and one in which only the most massive halos host sources (above $10^8\ M_\odot$, dashed curves). In both cases, we have normalized the mean amplitude of the Lyman-α background so that the background just reaches the coupling threshold, $\beta_\alpha \approx 1$, by setting $f_\star = 0.1$ and $z = 20$ and $z = 30$ for the high and low mass case, respectively. In both cases, intensity fluctuations are small on large scales: $< 1\%$ for $k < 0.1$ Mpc^{-1}, comparable to the horizon of each source (the wiggles in the power spectrum originate from the sharp cutoffs in the horizons of Lyman-n photons). If sources are common, the fluctuations remain small at smaller scales as

well, thanks to the enormous number of them. But if sources are rare, the "one-halo" term representing the intensity profile of each source becomes important at moderate or large scales, and the fluctuations can be moderately large on these scales. The dashed curve illustrates this: it assumes that each halo can only form stars in a single burst lasting 10 Myr (such as if each halo hosts a single short burst of Population III star formation). In this case, the fluctuations are much stronger, but the average radiation background also falls well below threshold.

In §11.5 we will consider how fluctuations in this background translate into fluctuations in the 21-cm signal.

11.3.2 Feedback: IGM Heating

The Wouthuysen-Field background couples the spin temperature to the gas kinetic temperature, so we must also compute the latter. A number of processes may contribute to it: shock heating from structure formation, ultraviolet photons, and X-rays.

The role of shock heating is unclear: very little of the IGM gas has been incorporated into sheets or filaments at these times, so the usual shocks that surround the cosmic web are unimportant. However, the very low temperature gas ($T < 30$ K) has large peculiar velocities from gravitational infall, and shocks may occur earlier in such an environment. We have seen that ionizing photons very effectively heat the gas, but the ionized gas that remains of course does not contribute to the 21-cm signal.[v]

The photons that trigger Lyman-α coupling do exchange energy with the IGM, through recoil. The typical energy exchange per scattering is small (see eq. 11.21), but the number of scatterings is large. If the net heating rate per atom were $\sim P_\alpha \times (h\nu_\alpha)^2/m_p c^2$, the gas temperature would exceed T_γ soon after Wouthuysen-Field coupling becomes efficient.

However, the details of radiative transfer radically change these expectations. In a static medium, the energy exchange must vanish in equilibrium even though scattering continues at nearly the same rate. Scattering induces an asymmetric absorption feature near ν_α (Fig. 11.6) whose shape depends on the combined effects of atomic recoils and the scattering diffusivity. In equilibrium, the latter exactly counterbalances the former. Without scattering, the absorption feature would redshift away; thus, the equilibrium energy exchange rate is simply that required to maintain the feature in place. For photons redshifting into resonance, the absorption trough has total energy $\Delta u_\alpha = (4\pi/c)\int (J_\infty - J_\nu) h\nu d\nu$, where J_∞ is the input spectrum (thus, the integration extends over the dip in Fig. 11.6). The radiation background loses $\epsilon_\alpha = H\Delta u_\alpha$ per unit time through redshifting; this energy goes into heating the gas. Relative to adiabatic cooling by the Hubble expansion,

[v] Moreover, the ionized regions will not significantly recombine unless somehow the source emissivity declines dramatically. The best example would be a highly luminous quasar that ionizes a large region around itself and then shuts off shortly thereafter.

the fractional heating amplitude is

$$\frac{2}{3}\frac{\epsilon_\alpha}{k_B T_K n_H H(z)} = \frac{8\pi}{3}\frac{h\nu_\alpha}{k_B T_K}\frac{J_\infty \Delta\nu_D}{c n_H}\int_{-\infty}^{\infty} dx \delta_J(x) \tag{11.34}$$

$$\approx \frac{0.80}{T_K^{4/3}}\frac{x_\alpha}{S_\alpha}\left(\frac{10}{1+z}\right), \tag{11.35}$$

Here we have evaluated the integral for the continuum photons that redshift into the Lyman-α resonance; the "injected" photons actually cool the gas slightly. The net energy exchange when Wouthuysen-Field coupling becomes important (at $x_\alpha \sim S_\alpha$) is therefore just a fraction of a degree, and this mechanism is usually unimportant.

The reason for the inefficiency of heating is that the scattering diffusivity acts to cancel the effects of recoil. From Figure 11.6, it is obvious that the background spectrum is weaker on the blue side of the line than on the red. Scattering tends to return the photon toward line center, with the extra energy deposited in or extracted from the gas. Because more scattering occurs on the red side, this tends to transfer energy from the gas back to the photons, canceling the recoil exchange.

Thus, IGM heating is likely dominated by X-rays – whether from Population III stars, supernova remnants, stellar-mass black holes, or quasars. We have already seen in equation (8.52) that X-rays from a "reasonable" quasar population can have a dramatic effect on the IGM, but even the weaker X-ray emissivity of stellar-mass black holes can also be significant.

A simple, but plausible, way to parameterize this emissivity is with the local correlation between the star formation rate (SFR) and the X-ray luminosity in the photon energy band of 0.5–8 keV,

$$L_X = 3\times 10^{39} f_X \left(\frac{\mathrm{SFR}}{M_\odot\ \mathrm{yr}^{-1}}\right)\ \mathrm{erg\ s}^{-1}, \tag{11.36}$$

where f_X is an unknown renormalization factor appropriate for high redshifts. We can only speculate as to the accuracy of this correlation at higher redshifts. Certainly the scaling is appropriate so long as recently-formed remnants 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 $u_{\rm CMB} \propto (1+z)^4$. Assuming that $\sim 5\%$ of the supernova energy is released in this form yields $f_X \sim 5$ if $\sim 10^{51}$ ergs are released in supernovae per 100 $M_\odot$in star formation. 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. To the extent that massive stars are more abundant at high redshifts (see the discussions of the IMF in chapter 5), we would expect such binaries to also be more abundant, which is consistent with some observational hints of evolution in this relation toward higher redshifts.

Regardless of the details of the sources, the heating rate and temperature profile around each source can be computed following the methods in §8.9.2. Note that, unlike the Wouthuysen-Field background, the IGM temperature depends not on the *instantaneous* emissivity of sources but on the accumulated emissivity over the entire history of structure formation. Thus the IGM temperature structure is more complicated to compute, although the same basic picture – built from the effects of each source halo – applies.

Figure 11.8 shows the temperature histories and power spectra of T_K for two models in which the heating is due to star-forming galaxies. In the left panels, the thick lines take $f_X = 10$ and standard Population II stars, forming with an efficiency $f_\star = 0.1$ in halos with $T_{\rm vir} > 10^4$ K; the thin lines are identical but take $f_\star = 0.01$ and use very massive Population III stars to determine the UV properties. Note that, even with this relatively modest heating rate, heating begins at $z \sim 15$ and the IGM temperature surpasses T_γ shortly thereafter. The right panels show the corresponding temperature power spectra; the top and bottom panels are for the Population III and Population II models, respectively. In absolute terms, the temperature fluctuations begin quite modestly; at $z = 20$ they are driven primarily by variations in the adiabatic cooling rate with IGM density. By $z = 15$, the fractional fluctuations are $\sim 20\%$ – which will translate into large 21-cm fluctuations. The absolute amplitude of the fluctuations continue to increase at lower redshifts, but the fractional fluctuations decrease as more sources appear; moreover, in the limit $T_K \gg T_\gamma$, the 21-cm brightness temperature is independent of T_K, so these fluctuations are unimportant.

11.4 THE MONOPOLE OF THE BRIGHTNESS TEMPERATURE

We are now in a position to compute the time evolution of the brightness temperature T_b in some simple models. We will begin in this section with the monopole, or sky-averaged brightness as a function of frequency. Figure 11.9 shows the results (as functions of both redshift and observed frequency) for a range of models, illustrating the wide range of possible histories. At left we show some highly simplified models. The solid curve with several turning points is our fiducial model, in which we take $f_X = 1$ and standard Population II stars, forming with an efficiency $f_\star = 0.1$ in halos with $T_{\rm vir} > 10^4$ K. The other two solid lines shows a history with no star formation (flat below $z \sim 30$) and with a hot, fully-coupled IGM (descending from large T_b). The dashed curve shows a history in which reionization does not occur, and the dotted curve shows a history in which heating is turned off.

The right panels take somewhat more sophisticated models, in which we vary the X-ray heating efficiency (via f_X, see eq. 11.36) and Lyman-α intensity (via a parameter f_α, defined so that the intensity from each galaxy is f_α times that in the fiducial model) by factors of 10^4.

These different models are essentially cartoons, but they illustrate several important points about the 21-cm background. The most important is the presence of five critical points in the spin-flip background.

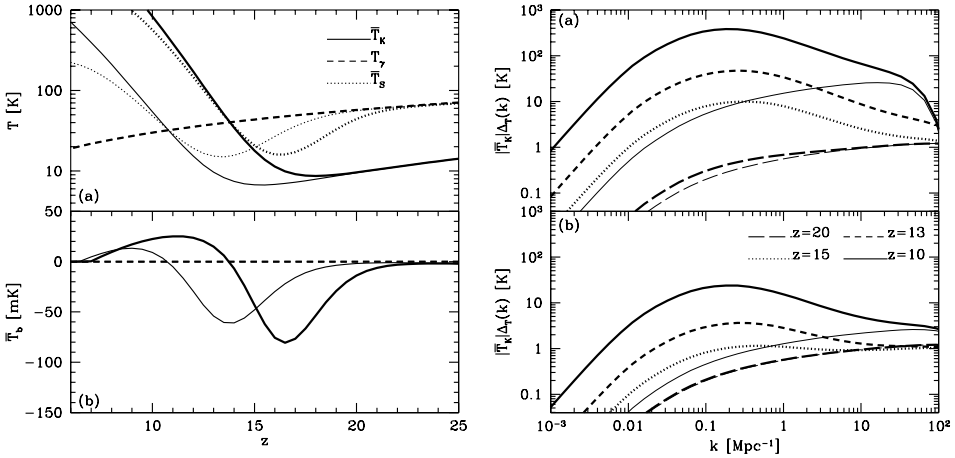

Figure 11.8 *Left:* Thermal history of two models of X-ray heating, Wouthuysen-Field coupling, and reionization. In the top panel, the solid, dashed, and dotted lines show T_K, T_γ, and T_S, respectively. The thick and thin lines take Population II and III star formation properties (see text for details). *Right:* Power spectrum of temperature fluctuations in the same models, from $z = 20$ (where the mean temperature is nearly that of an adiabatically cooling IGM) to $z = 10$ (where $T_K \gg T_\gamma$). The upper and lower panels are for the Population II and Population III models shown at left, respectively. The peak at late times comes from the typical mean free path of X-ray photons. The thin curves show the fluctuations for uniform heating for comparison. Figure credit: Pritchard & Furlanetto 2007, MNRAS, 376, 1680.

1. The first, at $z \sim 80$, occurs long before star formation becomes significant. This reflects the decreasing effectiveness of collisional coupling and occurs roughly when the density falls below $\delta_{\rm coll}$ (see Eq. 11.15), at which point $T_S \to T_\gamma$ and the IGM signal fades. This transition is well-specified by atomic physics and the standard cosmology, at least in the absence of any exotic dark sector processes that may input energy into the IGM at $z > 50$. This signal therefore provides a clear probe of cosmology, at least in priniciple.

2. The remaining transition points are determined by luminous sources, so their timing is much more uncertain. In our fiducial model, the next crucial event is the formation of the first stars (at $z \sim 30$), which flood the Universe with Lyman-α photons and so re-ignite the 21-cm background. Interestingly, the timing of this transition is relatively independent of the luminosity of these sources, because (at least in this model) the massive halos that host these sources are so far out on the exponential tail of the mass function that their luminosity is primarily determined by the rate of halo collapse. Thus, this turning point primarily constrains the characteristic mass of the first galaxies.

3. Next is the minimum in T_b, which occurs just before IGM heating begins

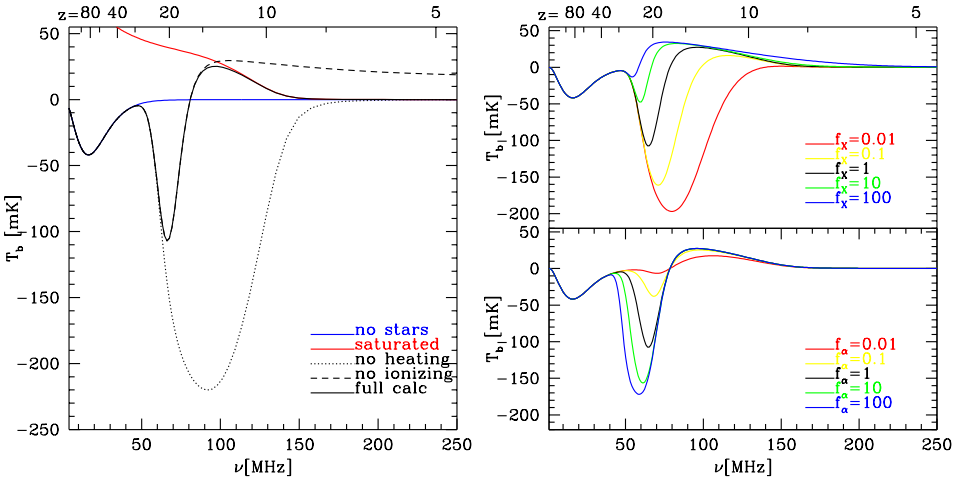

Figure 11.9 *Left:* Major variations around our fiducial model (solid curve with several turning points), as indicated. Each curve either eliminates a physical process (like heating or ionization) or maximizes it. *Right:* Suites of models in which we vary the Lyman-α (lower panel) and X-ray heating (upper panel) efficiencies by a factor of 10^4. Figure credit: J. R. Pritchard & A. Loeb, *Phys. Rev.***D82**, 023006 (2010).

to become significant and is determined by f_X. However, if this is very large compared to f_α, this heating transition can precede strong coupling. In simple models like we use here, in which both the X-ray and UV luminosities trace $f_{\rm coll}$, the net X-ray heat input ΔT_c when $x_\alpha = 1$ is

$$\frac{\Delta T_c}{T_\gamma} \sim 0.08 f_X \left(\frac{f_{X,h}}{0.2}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{9690}{N_\alpha}\frac{1}{S_\alpha}\right)\left(\frac{20}{1+z}\right)^3, \qquad (11.37)$$

where $\Delta f_{\rm coll} \sim f_{\rm coll}$ is the effective collapse fraction appearing in the integrals of equation (11.30). Note that ΔT_c is independent of $f_\star$ because both we have assumed that the coupling and heating rates are proportional to the star formation rate. Clearly, for our fiducial (Pop II) parameters the onset of Wouthuysen-Field coupling precedes the point at which $T_c \sim T_h$, which is ultimately responsible for the strong absorption in our fiducial model.

4. The third turning point is at the maximum of T_b. In the fiducial model, this marks the point at which $T_K \gg T_\gamma$, so that the temperature part of equation (11.8) saturates. From that point, the only factors affecting the monopole are the redshift and the ionized fraction, so the signal starts to decrease rapidly once reionization begins in earnest. Most likely, this happens *after* coupling is already strong and heating is significant. Again, in the simple models used here the ionized fraction when $x_\alpha = 1$ is given by

$$\bar{x}_{i,c} \sim 0.05 \left(\frac{f_{\rm esc}}{1+\bar{n}_{\rm rec}}\frac{N_{\rm ion}}{N_\alpha}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{1}{S_\alpha}\right)\left(\frac{20}{1+z}\right)^2, \qquad (11.38)$$

where $\bar{n}_{\rm rec}$ is the mean number of recombinations per baryon. For Population II stars with a normal IMF, $N_{\rm ion}/N_\alpha \approx 0.4$; 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 Population 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* $x_\alpha = 1$, although we have argued that this is rather unlikely.

It is less clear whether the IGM will appear absorption or emission during reionization. We find

$$\frac{\Delta T}{T_\gamma} \sim \left(\frac{\bar{x}_i}{0.025}\right) \left(f_X \, \frac{f_{X,h}}{f_{\rm esc}} \, \frac{4800}{N_{\rm ion}} \, \frac{10}{1+z}\right) (1+\bar{n}_{\rm rec}) \tag{11.39}$$

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. But the right panel of Figure 11.9 shows that this is by no means assured.

5. The monopole signal (nearly) vanishes when reionization completes; the residual brightness is due to gas that is self-shielded from the metagalactic ionizing background (and hence primarily lies inside of galaxies, since the LLSs still have small ionized fractions).

Several efforts to observe this monopole signal are underway, including the *Cosmological Reionization Experiment* (CoRE) and the *Experiment to Detect the Global Epoch of Reionization Signal* (EDGES)[vi]. The wide range of histories shown in Figure 11.9 illustrate how powerful such observations would be.

Because global experiments aim to detect an all-sky signal, single-dish measurements (even with a modest-sized telescope) can easily reach the required mK sensitivity. However, the much stronger synchrotron foregrounds from our Galaxy nevertheless make such observations extremely difficult: they have $T_{\rm sky} > 200$-10^4 K over the relevant frequencies (see the map in Figure 11.10). The fundamental strategy for extracting the cosmological signal relies on the expected spectral smoothness of the foregrounds (which primarily have power law synchrotron spectra), in contrast to the non-trivial structure of the 21 cm background. Nevertheless, extracting the high-redshift component will be a challenge that requires extremely accurate calibration over a wide frequency range and, most likely, sharp localized features in $\bar{T}_b(z)$ that can be distinguished from smoother foreground features.

Current estimates show that rapid reionization histories which span a redshift range $\Delta z < 2$ can be constrained, provided that local foregrounds can be well modeled. Observations in the frequency range 50-100 MHz can potentially constrain the Lyman-α and X-ray emissivity of the first stars and black holes: even though the foregrounds are significantly worse at these lower frequencies, the strong absorption signal present in many models may be easier to observe than the gently-varying reionization signal. However, it may be necessary to perform such observations from space, in order to avoid systematics from terrestrial interference and the ionosphere (in fact the best observing environment is the far side of the moon, where the moon itself blocks any radio signals from Earth).

[vi] See http://www.haystack.mit.edu/ast/arrays/Edges/.

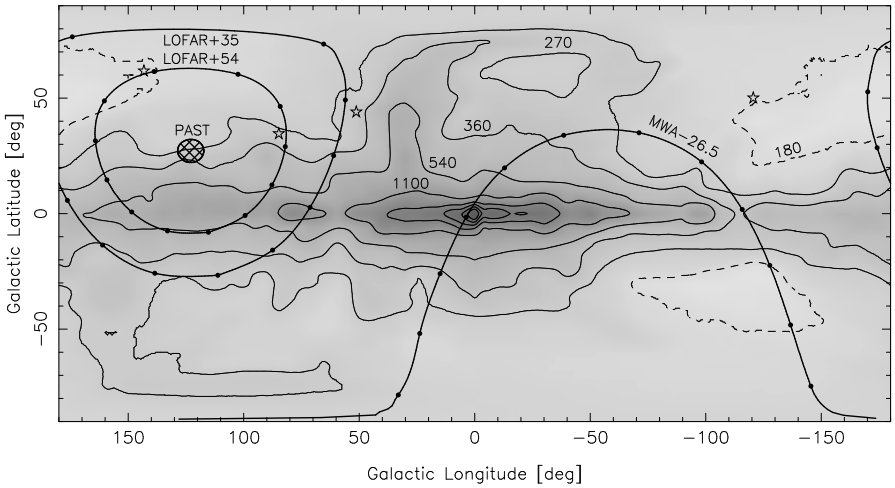

Figure 11.10 Brightness temperature of the radio sky at 150 MHz (from Landecker 1969) in Galactic coordinates. Contours are drawn at 180 (dashed), 270, 360, 540, 1100, 2200, 3300, 4400, and 5500 K. The 21-cm Array survey field at the North celestial pole is cross-hatched. Heavy lines indicate constant declinations:-26.5°, $+35^\circ$, and $+54^\circ$ with dots to mark 2 hour intervals of time (these are ideal for two other existing experiments, the Murchison Widefield Array (MWA) and LOFAR). Star symbols indicate the coordinates of four bright $z > 6.2$ quasars. Figure credit: S. R. Furlanetto, S.P. Oh, & F.H. Briggs, *Physics Reports* **433**, 181 (2006).

11.5 STATISTICAL FLUCTUATIONS IN THE SPIN-FLIP BACKGROUND

While the 21 cm monopole contains a great deal of information about the mean evolution of the sources, every component in equation (11.8) can also fluctuate significantly. For the density field this is obvious: the evolving cosmic web imprints growing density fluctuations on the matter distribution. For the other aspects, the discrete nature of the luminous sources gives rise to 21-cm fluctuations. Ionized gas is organized into discrete H II regions (at least in the most plausible models), and the Lyman-α background and X-ray heating will also be concentrated around galaxies. The single greatest advantage of the 21-cm line is that it allows us to separate this fluctuating component both on the sky and in frequency (and hence cosmic time). Thus, we can study the sources and their effects on the IGM in detail. It is the promise of these "tomographic" observations that makes the 21 cm line such a singularly attractive probe.

Observing the 21 cm fluctuations has one practical advantage as well. The difficulty of extracting the global evolution lies in its relatively slow variation with frequency. On the small scales relevant to fluctuations in the signal, the gradients increase dramatically: for example, at the edge of an H II region T_b drops by $\sim$ 20 mK essentially instantaneously. As a result, separating them from the smoothly varying astronomical foregrounds may be much easier. Unfortunately,

constructing detailed images will remain extremely difficult because of their extraordinary faintness; telescope noise is comparable to or exceeds the signal except on rather large scales. Thus, a great deal of attention has recently focused on using statistical quantities readily extractable from low signal-to-noise maps to constrain the IGM properties. This is motivated in part by the success of CMB measurements and galaxy surveys at constraining cosmological parameters through the power spectrum. In our case, although any number of statistical quantities may be useful (especially during reionization, when the fluctuations are highly non-gaussian), we will take the power spectrum as our primary analysis tool.

We first define the fractional perturbation to the brightness temperature, $\delta_{21}(\mathbf{x}) \equiv [T_b(\mathbf{x}) - \bar{T}_b]/\bar{T}_b$, a zero-mean random field. We will be interested in its Fourier transform $\tilde{\delta}_{21}(\mathbf{k})$. Its power spectrum is defined to be

$$\left\langle \tilde{\delta}_{21}(\mathbf{k}_1)\, \tilde{\delta}_{21}(\mathbf{k}_2) \right\rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}_1 - \mathbf{k}_2) P_{21}(\mathbf{k}_1), \tag{11.40}$$

where $\delta_D(x)$ is the Dirac delta function and the angular brackets denote an ensemble average. Power spectra for other random fields (such as the fractional overdensity δ, the ionized fraction, etc.), or cross-power spectra between two different fields, can be defined in an analogous fashion.

As is obvious from equations (11.8) and (11.10), the brightness temperature depends on a number of input parameters. Expanding those equations to linear order in each of the perturbations, we can write

$$\delta_{21} = \beta\delta_b + \beta_x\delta_x + \beta_\alpha\delta_\alpha + \beta_T\delta_T - \delta_{\partial v}, \tag{11.41}$$

where each δ_i describes the fractional variation in a particular quantity: δ_b for the baryonic density (for which we will use the matter density, though the baryonic density is smoother on very small scales thanks to pressure smoothing), δ_α for the Lyman-α coupling coefficient x_α, δ_x for the neutral fraction (note that using the ionized fraction would cause a sign change), δ_T for T_K, and $\delta_{\partial v}$ for the line-of-sight peculiar velocity gradient. The expansion coefficients β_i are

$$\beta = 1 + \frac{x_c}{x_{\rm tot}(1 + x_{\rm tot})}, \tag{11.42}$$

$$\beta_x = 1 + \frac{x_c^{\rm HH} - x_c^{\rm eH}}{x_{\rm tot}(1 + x_{\rm tot})} \tag{11.43}$$

$$\beta_\alpha = \frac{x_\alpha}{x_{\rm tot}(1 + x_{\rm tot})}, \tag{11.44}$$

$$\beta_T = \frac{T_\gamma}{T_K - T_\gamma} + \frac{1}{x_{\rm tot}(1 + x_{\rm tot})} \left(x_c^{\rm eH} \frac{d\ln\kappa_{10}^{\rm eH}}{d\ln T_K} + x_c^{\rm HH} \frac{d\ln\kappa_{10}^{\rm HH}}{d\ln T_K} \right), \tag{11.45}$$

where $x_{\rm tot} \equiv x_c + x_\alpha$ and we have split the collisional term into the dominant H-e$^-$ and H-H components ($x_c^{\rm eH}$ and $x_c^{\rm HH}$, respectively) where necessary. Here we have assumed $T_c = T_K$ throughout; this is reasonable in most cases but, if not, the expressions become much more complicated. By linearity, the Fourier transform $\tilde{\delta}_{21}$ can be written in a similar fashion.

Each of these expressions has a simple physical interpretation. For β, the first term describes the increased matter content and the second describes the increased

collisional coupling efficiency in dense gas. For β_x, the two terms describe direct fluctuations in the ionized fraction and the effects of the increased electron density on x_c. (The latter is only important in partially ionized regions; 21 cm emission is negligible in HII regions, of course.) β_α simply measures the fractional contribution of the Wouthuysen-Field effect to the coupling. The first term in β_T parameterizes the speed at which the spin temperature responds to fluctuations in T_K, while the others include the explicit temperature dependence of the collision rates. Note that all of these terms, with the crucial exception of $\delta_{\partial v}$, are isotropic; we will discuss the velocity term in the next section.

For context, Figure 11.11 shows how these expansion coefficients evolve in a typical structure formation model (similar to those described in the previous section). The density coefficient β increases with time until $z \sim 20$ before abruptly falling to unity. At $z > 20$, collisions are only marginally important so the extra collisional coupling imparted by an increased density has a relatively large effect; at lower redshifts, collisional coupling is negligible compared to the Wouthuysen-Field effect so the second term in equation (11.42) vanishes. β_x behaves nearly identically, because (outside of H II regions) the ionized fraction remains small. Fluctuations in the Lyman-α background are only important over a limited redshift range (where $x_\alpha \sim 1$, or the coupling is marginal); at lower redshifts, all the gas is strongly coupled so fluctuations in the background are unimportant. The temperature coefficient has the most complicated dependence because it depends on the mix of Compton heating and collisional coupling. Note that the apparent singularity occurs where $T_K = T_\gamma$; it is not physical because $\bar{T}_b$ also vanishes at the same point. At lower redshifts, $T_K \gg T_\gamma$ and the emission saturates, $\beta_T \to 0$.

Based on equation (11.40), the power spectrum contains all possible terms of the form $P_{\delta_i \delta_j}$; some or all could be relevant in any given situation. Of course, in most instances the various δ_i will be correlated in some way; statistical 21 cm observations ideally hope to measure these separate quantities. We have already included some of the obvious correlations in equations (11.42)–(11.45), such as the variation of the collision rate with the ionized fraction. But we have left others implicit: for example, overdense regions are ionized first in most reionization models. A more subtle example is the relation of δ_α to the other quantities; as we saw in §11.2.2, it depends on the radiation spectrum and hence on density, neutral fraction, and temperature in addition to the background flux.

In all of these expansions, one must bear in mind that δ_x is of order unity if the ionization field is built from H II regions. In that case terms such as $\delta\delta_x$ are in fact *first* order and must be retained in detailed calculations.

11.5.1 Redshift-Space Distortions

In general, we expect the fluctuations in density, ionization fraction, Lyα flux, and temperature to be statistically isotropic, because the physical processes responsible for them have no preferred direction [e.g., $\delta(\mathbf{k}) = \delta(k)$].[vii] However, peculiar ve-

[vii] Actually, this assumption can break down on extremely large scales, because then the growth of structure with redshift becomes important. Fortunately, the 21-cm brightness field only contains rapidly evolving features on such large scales at the tail end of reionization. The evolution is generally not

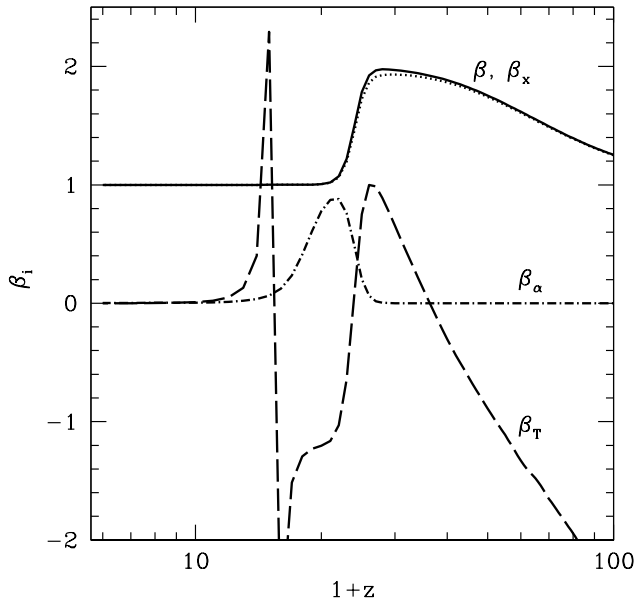

Figure 11.11 Redshift dependence of perturbative expansion coefficients in a fiducial model similar to that of Fig. 11.9. We show β (solid curve), β_x (dotted curve), β_α (dot-dashed curve), and β_T (dashed curve). Note that the singularity in β_T at $z = 17$ is artificial in that it does not actually appear in the fluctuation amplitude. Figure credit: S.R. Furlanetto, S.P. Oh, F.H. & Briggs, *Physics Reports* **433**, 181 (2006).

locity gradients introduce anisotropic distortions. Bulk flows on large scales, and in particular infall onto massive structures, compress the signal in redshift space (the so-called "Kaiser" effect [41]), enhancing the apparent clustering amplitude. On small-scales, random motions in virialized regions create elongation in redshift space (the "finger of God" effect), reducing the apparent clustering amplitude.

We start by labeling the coordinates in redshift space by $\mathbf{s}$. Working for simplicity under the assumption that the survey volume has a small radial depth (so that the Hubble parameter H can be considered constant throughout the volume), these coordinates are related to real space by

$$\mathbf{s}(\mathbf{r}) = \mathbf{r} + \frac{U(\mathbf{r})}{H}, \tag{11.46}$$

where $U(r) = \mathbf{v} \cdot \hat{\mathbf{x}}$ is the radial component of the peculiar velocity.

Next, we consider a set of particles with number density $n(\mathbf{r})$ that are biased with respect to the dark matter by a factor b. Number conservation demands that the fractional overdensity in redshift space is related to that in real space via $[1 +$

important on the scales accessible to observations.

$\delta_s(\mathbf{s})]d^3\mathbf{s} = [1+\delta(\mathbf{r})]d^3\mathbf{r}$. The Jacobian of the transformation is

$$d^3\mathbf{s} = d^3\mathbf{r}\left[1+\frac{U(\mathbf{r})}{r}\right]^2\left[1+\frac{dU(\mathbf{r})}{dr}\right], \tag{11.47}$$

because only the radial component of the volume element, r^2dr, changes from real to redshift space. Thus, the density observed in redshift space increases if the peculiar velocity gradient is smaller than the Hubble flow or decreases otherwise. Thus, assuming $|U(r)| \ll Hr$,

$$\delta_s(\mathbf{r}) = \delta(\mathbf{r}) - \left(\frac{d}{dr}+\frac{2}{r}\right)\frac{U(r)}{H}. \tag{11.48}$$

Conveniently, the peculiar velocity field itself is a function of the dark matter density field, as described by equation (2.10).

To see which of these corrections is more important, consider a plane wave perturbation, $U \propto e^{i\mathbf{k}\cdot\mathbf{r}}$. Then the derivative term is $\sim kU/H_0$ while the last term is $\sim U/H_0r$. But r is the median distance to the survey volume, and k corresponds to a mode entirely contained inside it. For all but the largest surveys, we must therefore have $kr \gg 1$, and we may neglect the last term. If we further make the small-angle approximation, so that $\hat{\mathbf{x}}$ is also approximately a constant over the survey volume, we can take the Fourier transform of equation (11.48) and find

$$\delta_s(\mathbf{k}) = \delta(\mathbf{k})[1+\beta\mu_{\mathbf{k}}^2] \tag{11.49}$$

where $\mu_{\mathbf{k}} = \hat{\mathbf{k}}\cdot\hat{\mathbf{x}}$ is the cosine of the angle between the wave vector and the line of sight, and we have used

$$U(r) = \int \frac{d^3k}{(2\pi)^3}e^{i\mathbf{k}\cdot\mathbf{x}}[-i\beta\delta(\mathbf{k})]\frac{\hat{\mathbf{k}}\cdot\hat{\mathbf{x}}}{k}. \tag{11.50}$$

Here $\beta = f(\Omega_m)/b$ corrects for a possible bias between the tracers we are studying and the growth rate of dark matter perturbations, and $f(\Omega_) \approx \Omega_m^{0.6}(z)$. For the case of 21-cm fluctuations in the IGM gas, the bias factor is very close to unity except below the filtering scale. Moreover, at high-redshifts when the universe is matter-dominated, $f \approx 1$.

The redshift-space distortions therefore provide an anisotropic *amplification* to the background signal. The anisotropy occurs because only modes along the line of sight are effected. To understand the amplification, consider a spherical overdense region. Its excess gravitational force causes it to recollapse. Along the radial direction, the collapse *decreases* the velocity width of the object relative to the Hubble flow (at least in linear theory), compressing the overdensity in redshift space. Similarly, a spherical underdensity expands faster than average, causing it to appear elongated in the radial direction. Averaged over all modes, these distortions amplify the signal by a factor $\approx \left\langle(1+\mu^2)^2\right\rangle \approx 1.87$.[viii]

[viii] In most applications, there is a second signature of peculiar velocities that tends to wash out fluctuations in redshift space: the "fingers of God" which cause an apparent smearing along the line of sight due to random velocities inside virialized structures. Fortunately, this effect is negligible for the spin-flip background because the vast majority of the gas lies outside of massive virialized structures (and gas inside such halos is almost always inside ionized regions anyway).

However, the anisotropies are even more helpful in that they provide angular structure to the signal, which may allow us to separate the many contributions to the total power spectrum. Schematically, brightness temperature fluctuations in Fourier space have the form

$$\delta_{21} = \mu^2 \beta \delta + \delta_{\rm iso} \tag{11.51}$$

where we have collected all the statistically isotropic terms in equation (11.41) into $\delta_{\rm iso}$. Neglecting "second-order" terms (see below) and setting $\beta = 1$, the total power spectrum can therefore be written as

$$P_{21}(\mathbf{k}) = \mu^4 P_{\delta\delta} + 2\mu^2 P_{\delta_{\rm iso}\delta} + P_{\delta_{\rm iso}\delta_{\rm iso}}. \tag{11.52}$$

By separately measuring these three angular components (which requires, in principle, estimates at just a few values of μ), we can isolate the contribution from density fluctuations $P_{\delta\delta}$. This would not have been possible without peculiar velocity flows: comparison to equation (11.41) shows that, in the most general case, $P_{\delta_{\rm iso}\delta}$ and $P_{\delta_{\rm iso}\delta_{\rm iso}}$ contain several different power spectra, including those of the density, neutral fraction, and spin temperature as well as their cross power spectra.

Disentangling these other components is more difficult, since there are several remaining power spectra to be determined from the two measured quantities $P_{\delta_{\rm iso}\delta}(k)$ and $P_{\delta_{\rm iso}\delta_{\rm iso}}(k)$. Fortunately, in many regimes one or more of the terms can be neglected. For example, during the earliest stages of reionization (when δ_x is negligible), one might be able to measure the power spectrum of spin temperature fluctuations as well as its correlations with density. At late times (when $T_S \gg T_\gamma$ and T_b becomes independent of T_S), one might likewise ignore spin temperature fluctuations and measure the ionization fraction fluctuations $P_{\delta x}$ and P_{xx}.

An additional difficulty originates from the correlations of "second-order" terms in the perturbation expansion, such as $\delta\delta_x$, that produce four-point terms in the power spectrum. Unfortunately, δ_x is not necessarily a small parameter, so these terms can be substantial, and in practice they can produce terms with non-trivial μ dependence, especially during reionization. The presence of these terms make attempts to separate the μ^n powers during reionization more difficult; the prospects are much better before δ_x becomes important.

Another important caveat to recovering redshift space distortions is that they require a high signal to noise measurement of the angular structure of the signal. Unfortunately, the noise is anisotropic: radio foregrounds have much more power across the sky than in the line of sight direction. (Indeed, this very feature is crucial to foreground removal algorithms.) Moreover, it is much easier to probe small physical scales in the frequency direction than across the angular dimensions. As a result, taking advantage of this "separation of powers" will be difficult.

11.5.2 Other Statistical Measures

So far we have focused our discussion on the three-dimensional power spectrum, which is familiar to most cosmologists and provides a reasonable description of the spin-flip background during most of its evolution. In fact the power spectrum is a complete statistical description of any purely Gaussian random field (whose only

parameters are, by definition, the mean and variance as a function of spatial scale). Inflation predicts that the initial matter density field is nearly Gaussian, making the power spectrum a powerful tool in cosmology.

However, nonlinear evolution – and the radiation fields from such sources – spoil this simple statistical description for the 21-cm fluctuations, especially when ionized bubbles become prevalent late in reionization. It is easy to see that a Gaussian probability distribution will no longer adequately describe the 21-cm field: at infinite resolution, the signal is either nearly zero (in an ionized bubble) or ~ 20 mK (in the neutral gas, where there is still some variation due to the density field and possibly T_S). This bivariate distribution is a strong signature of ionized bubbles and would provide a powerful test of the morphology of reionization; unfortunately, in experiments where the Gaussian noise per pixel is larger than ~ 20 mK this kind of distribution may be difficult to detect, especially in the presence of complex astrophysical foregrounds.

Other statistical measures, such as higher-order correlations, may also offer additional information and are the subject of ongoing research in the community.

11.6 SPIN-FLIP FLUCTUATIONS DURING THE COSMIC DAWN

Figure 11.12 shows several snapshots of a "semi-numeric" computer simulation (see §8.7.2) of the spin-flip background during the important stages outlined in our discussion of the monopole signal of §11.4, including both snapshots of the fields (in the left column) and the corresponding (spherically-averaged) power spectra (in the right column). The underlying model is very similar to the fiducial model whose mean signal is shown in Figure 11.9, though the redshifts of the critical points differ slightly. Importantly, the fluctuations are substantial throughout all of the interesting regimes.

The top row of Figure 11.12 shows the point where Lyman-α pumping begins to be significant. The hydrogen gas is cold ($T_K \ll T_\gamma$), and the spin temperature is just beginning to decouple from the CMB. In this case the fluctuations are driven by the discrete, clustered first galaxies: their fluctuating radiation field drives $T_S \rightarrow T_K$ around those first sources, while leaving most of the IGM transparent.

In this model, the Lyman-α radiation field builds up very quickly the brightness temperature fluctuations. We illustrate this in Figure 11.13, which shows the evolution of the amplitude of the power spectrum at one particular wavenumber ($k = 0.1$ Mpc^{-1}, near the peak sensitivities of most arrays).[ix] The dashed curve shows the effects of the Lyman-α fluctuations: they build up to a peak, with amplitude ~ 10 mK, before decreasing again once Lyman-α fluctuations become strong everywhere (so that $\beta_\alpha \propto 1/x_\alpha \rightarrow 0$).

The second row in Figure 11.12 shows the signal near the onset of X-ray heating. In this model, the Lyman-α coupling is strong nearly everywhere, so most of the IGM is cold and hence appears in absorption. But near the first X-ray sources (as-

[ix] This example is taken from a different analytic model, so the times at which the critical points differ relative to the semi-numeric calculation. However, the qualitative evolution is identical.

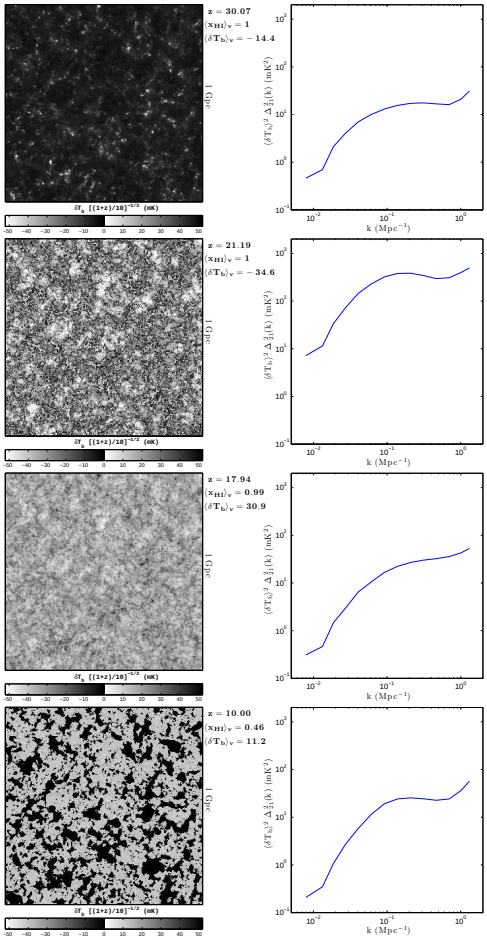

Figure 11.12 Slices through a "semi-numeric" simulation (*left*), and the corresponding spherically-averaged power spectra (*right*), for a model of the spin-flip background at $z = 30.1$, 21.2, 17.9, 10.0 (*top to bottom*). The slices were chosen to highlight various epochs in the cosmic 21-cm signal: the onset of Lyman-α pumping (here the white regions show the cold gas around the first galaxies), the onset of X-ray heating (here the larger white regions are typically cold gas, while the compact light and dark regions represent hot gas around the first black holes), the completion of X-ray heating (where all the gas is hot), and the mid-point of reionization (where black regions are ionized bubbles) are shown from top to bottom. All comoving slices are 1 Gpc on a side and 3.3 Mpc deep. Figure credit: A. Mesinger, S.R. Furlanetto, & R. Cen, *Mon. Not. R. Astron. Soc.* **411**, 955 (2011).

sumed to be star-forming galaxies here), the X-ray background has already heated the gas to $T_S \gg T_\gamma$, so these regions appear in emission. The net effect is a very large fluctuation amplitude, with a strong contrast between emitting and absorbing

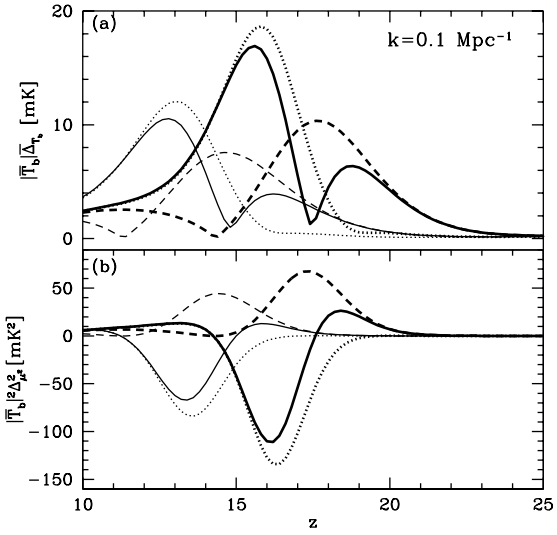

Figure 11.13 Evolution of brightness temperature fluctuations at $k = 0.1\,\mathrm{Mpc}^{-1}$ in two models of the early history of the 21-cm background; this is near the peak sensitivity of most of the planned experiments. The thick curves use parameters similar to our fiducial model in Figure 11.9 (and also the thick curves in Figure 11.8); the thin curves use the Population III model from Figure **??**. The dashed curve includes the effects of Lyman-α fluctuations *only*, the dotted curve includes the effects of heating fluctuations *only*; the solid curve includes both. Figure credit: J.R. Pritchard & S.R. Furlanetto, *Mon. Not. R. Astron. Soc.* **376**, 1680 (2007).

regions.

Figure 11.13 also helps to illustrate this behavior. Here the dotted curve includes *only* the effects of heating fluctuations (implicitly assuming strong Lyman-α coupling throughout). The signal rises to ~ 20 mK when this strong contrast is in place; then the fluctuations decrease once more of the IGM becomes hot (and hence saturates in emission).

The solid curve includes both heating and Lyman-α fluctuations. In this model the X-ray background lags the Wouthuysen-Field coupling, but not by a large margin. As a result, the net signal actually *decreases* in the early phases of the heating era. This occurs because only the regions near the first sources have strong coupling, but these are also the regions that are heated; the resulting emission signal is weaker than absorption because of the saturation in equation (11.8). Once the Lyman-α background reaches more of the IGM, the signal increases quickly.

The third row in Figure 11.12 shows the 21-cm signal once heating has saturated ($T_S \gg T_\gamma$) throughout the IGM. At this point, spin temperature fluctuations no longer contribute to T_b, and only the density field affects the overall signal. The fluctuations are thus relatively modest (as in the late stages of the model of Fig-

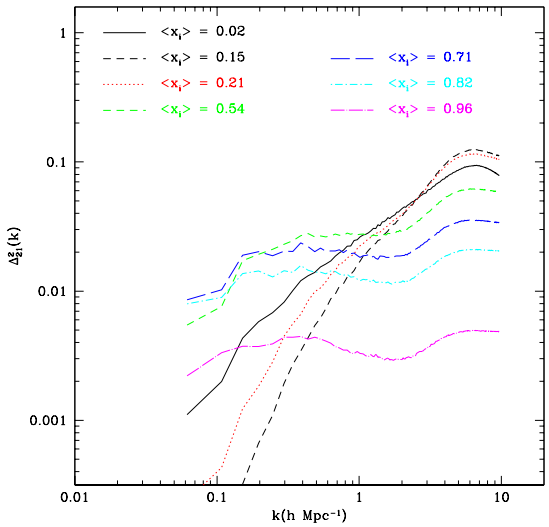

Figure 11.14 Dimensionless power spectra $\Delta_{21}(k)$ of spin-flip background during the reionization era in a numerical simulation with radiative transfer; to obtain the 21-cm signal one needs to multiply $\Delta_{21}(k)$ by the mean brightness temperature in a fully neutral medium, $\sim [28^2(1+z)/10]$ mK^2. The curves show the power spectrum through a sequence of mean ionized fractions; the redshifts at which these points are achieved (not listed) do not significantly affect the signal, except through the mean brightness temperature. Figure credit: Lidz, A. et al. 2008, ApJ, 680, 982.

ure 11.13). However, this period could be very important for cosmological measurements, because the astrophysical uncertainties in the ionized fraction and T_S are not significant (see below).

Finally, the fluctuations increase again once reionization begins in earnest, as shown in the bottom row of Figure 11.12: here the fluctuations in the map are dominated by the contrast between the ionized bubbles and fully neutral gas in between them. As we saw in §8.5, the pattern of these bubbles contains information about the ionizing sources creating them.

Figure 11.14 shows how the dimensionless power spectrum evolves in a radiative transfer simulation of the reionization process. (To recover the 21-cm signal one needs to multiply these values by the mean brightness temperature in a fully neutral medium, $T_0 \approx [28^2(1+z)/10]$ mK^2.) The different curves show a sequence of ionized fractions, from nearly neutral ($\langle x_i \rangle = 0.02$) to almost fully ionized ($\langle x_i \rangle = 0.96$). In this model, these span a range of redshifts from $z \sim 11.5$–6.8, but as we have shown earlier the curves change little if one holds $\langle x_i \rangle$ constant but chooses a different redshift.

Clearly the shape and amplitude of the power spectrum both evolve substantially throughout reionization. At first, the 21-cm power spectrum simply traces the mat-

ter power spectrum, as ionized regions have not yet significantly affected the IGM (and in this model $T_S \gg T_\gamma$ throughout the IGM, so spin temperature fluctuations are likewise unimportant). The power then *decreases* on large scales because the ionized bubbles appear first in the densest regions, suppressing the signal there and hence decreasing the overall contrast in the 21-cm maps.

This is simplest to understand if we decompose the power spectrum into parts that describe perturbations in each relevant physical parameter and retain only the dominant components (see eq. 11.41)

$$\Delta^2_{21}(k) = T_0^2 \langle x_H \rangle^2 \left[\Delta^2_{\delta\delta}(k) + 2\Delta^2_{x\delta}(k) + \Delta^2_{xx}(k)\right]. \quad (11.53)$$

In this equation, $\Delta^2_{21} = k^3 P_{21}(k)/2\pi^2$ is the power per logarithmic interval in wavenumber of the 21-cm signal, $\Delta^2_{\delta\delta}$ and Δ^2_{xx} represent the power spectra of the density field and ionized fraction, and $\Delta^2_{x\delta}$ is the cross-power spectrum of the density with the ionized fraction. Here we have included only "low-order" terms in which two quantities are correlated, for simplicity. In fact, because the ionized fraction is usually either ≈ 0 or ≈ 1, "higher-order" terms such as $\Delta^2_{x\delta,x\delta}$, expressing the joint correlations between ionized fraction and density evaluated at two different locations, is not necessarily smaller than the terms we have retained. Because $\Delta^2_{x\delta}$ is this cross-power, it can be negative – i.e., the neutral fraction x_H is small when δ is large in most reionization models. In the early phases of reionization, this term dominates the ionized power itself, Δ^2_{xx}, and so the net power falls.

However, by $\langle x_i \rangle \sim 0.5$, the ~ 20 mK contrast between ionized and neutral gas the maps dominates the maps, and the power increases rapidly: now the ionized bubbles fill a wide range of density, so $\Delta^2_{x\delta}$ is small but Δ^2_{xx} is large – at least on large scales. In fact the power from this term peaks on the characteristic scale of the ionized bubbles (which is well-defined in most reionization models; see Figure 8.3). In combination with the contribution from the matter power spectrum itself, this leads to a strong enhancement of power on moderate scales ($k \sim 0.1\,\mathrm{Mpc}^{-1}$), followed by a decline at smaller wavenumbers (not shown clearly in this figure because of the finite size of the simulation box).

At the same time, on scales much smaller than the bubble size, the 21-cm power is significantly smaller than expected from the matter power spectrum alone. This is largely because of the higher-order terms that we have ignored: within an ionized region, the ionized fraction is largely uncorrelated with the small-scale density perturbations. Effectively then the contrast on these scales is decreased because many of the small-scale overdensities no longer appear in the 21-cm map. The net effect is an overall *flattening* in Δ^2_{21} throughout reionization. The flattening shifts to larger scales throughout reionization, and the amplitude decreases as less of the gas can emit 21-cm photons.

Because the power spectrum of the ionized fraction dominates the signal on large scales, the spin-flip background could be an effective tool to study the morphology of reionization (and the sources that drive it); the shape and amplitude of the power spectrum can inform us of the time history of reionization throughout the IGM and (through the bubble size distribution) the clustering properties of the sources that drive it. This interpretation is relatively model-independent in contrast to galaxy surveys whose implications for reionization are difficult to interpret due to the many

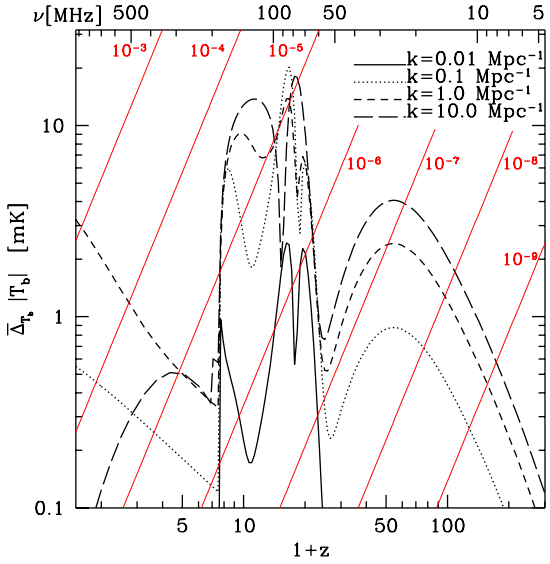

Figure 11.15 Redshift evolution of the angle-averaged 21 cm power spectrum in a model with reionization ending at $z = 6.5$. We show the amplitude for $k = 0.01$ Mpc^{-1} (solid curve), $k = 0.1$ Mpc^{-1} (dotted curve), $k = 1$ Mpc^{-1} (short-dashed curve), and $k = 10$ Mpc^{-1} (long-dashed curve). After reionization, the fluctuations trace neutral gas inside galaxies and DLAs and so mirror the galaxy power spectrum. The diagonal curves show contours of a fixed fraction of the sky brightness as a function of frequency. Figure credit: Pritchard, J. R. & Loeb, A. *Phys. Rev.* **D 78**, 103511 (2008).

unknown properties of the observed galaxies.

The final phase in the evolution of the 21-cm background is the end of reionization, when the vast majority of the gas is ionized and so the spin-flip signal declines dramatically. But it does not disappear: substantial reservoirs of neutral gas still exist inside of self-shielded galaxy-sized objects – the "damped Lyman-α absorbers" we have discussed before. Observations show that these systems typically have $T_S \gg T_\gamma$; in this limit the power spectrum is simply

$$\Delta^2_{21}(k) \approx T_0^2 \, \langle x_H \rangle^2 \, \Delta^2_{gg}(k), \tag{11.54}$$

where Δ^2_{gg} is the galaxy power spectrum (which can be computed easily with the halo model) and $\langle x_H \rangle$ is measured from a census of DLAs to be a few percent after reionization. Figure 11.15 compares the post-reionization signal to the higher redshift one at several different wavenumbers in a model where reionization is tuned to end at $z = 6.5$. Provided that galactic systems do dominate the neutral gas, fluctuations in the spin-flip background at redshifts after reionization therefore present an interesting cosmological probe – with the same information as galaxy surveys – but offer little information about the IGM itself.

11.6.1 Extracting Cosmological Measurements from the Spin-Flip Background

To this point, we have focused on the spin-flip background as a rich *astrophysical* data set. However, it also holds great promise for measurements of "fundamental" cosmological information, much like the CMB. There are several reasons for this promise. First, the 2-cm signal probes a time period when structure formation is still in its infancy – and, in particular, still within the well-understood linear regime through most of space. Second – unlike with galaxy surveys – the 21-cm signal probes the majority of baryonic matter that lies outside of virialized structures, allowing us to access directly the linear fluctuations in the matter field.

Third – unlike the CMB – a 21-cm survey yields a three-dimensional data set and hence probes a much larger fraction on the cosmic volume. We will see in the next section that the ultimate fractional uncertainty in the amplitude of any Fourier mode of wavelength λ is given by $\sim 1/\sqrt{N}$, where N is the number of independent elements of size λ that fit within the survey volume. For the two-dimensional map of the CMB, N is the surveyed area of the sky divided by the solid angle occupied by a patch of area λ^2 at $z \sim 10^3$. For a three-dimensional field, we obtain one of these maps at every frequency, vastly increasing the size of the available data set. Figure 11.16 shows the fraction of the total comoving volume of the observable Universe that is available up to different redshifts. Clearly 21-cm surveys at $z \sim 10$ probe a much bigger comoving volume than conventional galaxy surveys at $z < 1$.

Finally, the 21-cm power extends down to the pressure-dominated (Jeans) scale of the cosmic gas. This is orders of magnitude smaller than the comoving scale at which the CMB anisotropies are damped by photon diffusion. Consequently, the spin-flip background can trace the primordial inhomogeneities with a much finer resolution (i.e. many more independent pixels) than the CMB. Altogether, the above factors imply that 21-cm tomography of cosmic hydrogen may potentially carry more information about the initial conditions of our Universe than any other method.

Of course, extracting cosmological information in the presence of the rich astrophysics that sets the 21-cm brightness may be challenging. Fortunately, there are two regimes in which it may be possible. The first is *before* the first stars light up the Universe. During these "dark ages," there is no astrophysics that can possibly interfere; however, at such high redshifts, corresponding to low frequencies, the noise is extraordinarily large, so this era will remain inaccessible for the foreseeable future.

A second possibility is if the ionization and temperature factors in equation (11.8) can both be neglected; then the spin-flip brightness traces density and velocity fluctuations, both of which can be easily translated into fundamental cosmological parameters like the matter content or Hubble constant. We saw in §11.4 that such a scenario is plausible: the Wouthuysen-Field effect can become strong long before reionization begins in earnest, and X-ray heating could also be very fast – but this outcome is by no means guaranteed. If not, cosmological information can be extracted only if the astrophysics is well understood.

We will see later that high-precision measurements of the 21-cm background are challenging, and for the foreseeable future the direct constraints on, e.g., the

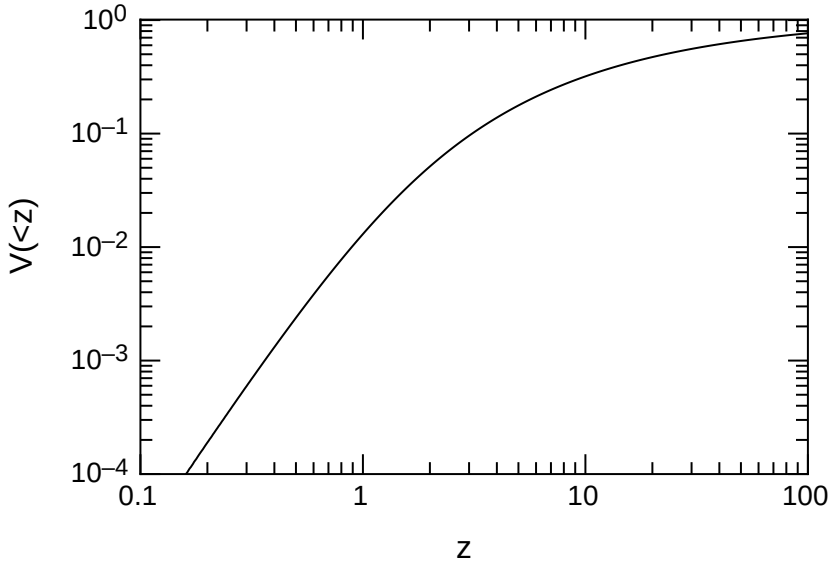

Figure 11.16 The fraction of the total comoving volume of the observable Universe that is available up to a redshift z, as a function of z. Image credit: A. Loeb, & J.S.B. Wyithe, *Phys. Rev. Lett.* **100**, 161301 (2008).

matter power spectrum will not be competitive with those from galaxy surveys or the CMB. However, because the spin-flip background extends to such small scales, it still adds new cosmological information compared to other measurements. This is particularly useful for cosmological parameters that depend crucially on small scales, such as the shape of the primordial power spectrum and the neutrino mass (because free-streaming of neutrinos erases small-scale power).

11.7 MAPPING THE SPIN-FLIP BACKGROUND

The prospect of studying reionization by mapping the distribution of atomic hydrogen across the Universe through its 21-cm spectral line has motivated several teams to design and construct arrays of low-frequency radio antennae. For redshifts $z \sim 6$–50, the corresponding observed frequencies are $\nu_{\rm obs} \sim 30$–200 MHz. Although the radio technology for the frequency range of interest has existed for decades – and is essentially the same that we use everyday for TV or radio communication – these experiments face three extreme challenges before they can observe the spin-flip background:

- The low-frequency band is heavily used by humans (as it includes the FM radio band, analog TV stations, and a host of satellite and aircraft communications channels), and the resulting **terrestrial radio interference** is as many as ten orders of magnitude brighter than the 21-cm background. Most of the efforts therefore place the observatories in isolated locations far from the contaminating sources (although some residual contamination does remain). However, these signals are usually (though not always) narrowband, so one can also attempt to measure the signal only in the gaps between contaminated channels. Even then, the presence of such bright foregrounds places serious requirements on the dynamic range of the low-frequency observatories.

- The **ionosphere** is refractive at low frequencies (and at the lowest frequencies, corresponding to redshifts $z > 50$, becomes opaque). This causes sources to jitter across the sky as patches of the ionosphere move across the telescope beam. The refraction phenomenon is similar to seeing in optical astronomy, although the timescale for the jitter is much slower (several seconds in this case). It can be corrected in software by calibrating to the locations of a set of point sources distributed across the field of view, although this is by no means a trivial computing effort. The ionosphere is more active during the day and during times of high solar activity. This – together with the large brightness of the sun itself at these frequencies – restricts these observatories to operate only at night.

- Most significantly, the spin-flip background is far from the only astronomical source in the sky. Nearly all non-thermal radio sources are bright in the low-frequency band, especially the synchrotron radiation from the Milky Way galaxy, as we have already seen in Figure 11.10. But other extragalactic sources – including AGN, galaxy clusters, and even normal star-forming

galaxies – also contribute. A rule of thumb, typical high-latitude, "quiet" portions of the sky have

$$T_{\rm sky} \approx 180 \left(\frac{\nu}{180\,{\rm MHz}}\right)^{-2.6} {\rm K}. \quad (11.55)$$

This brightness is so large that it swamps the noise from even a simple receiver, so the net system temperature of one of these instruments can be well-approximated by $T_{\rm sys} \approx T_{\rm sky}$. We immediately see that 21-cm mapping will require large integration times and large collecting area to overcome this "noise,", which is at least 10^4 times stronger than the cosmic signal.

Despite numerous efforts over the past four decades to observe the spin-flip background, these factors – as well as not-yet mature theories of the first galaxies – conspired to prevent any detection. Now, with modern computing, it has become possible to analyze the enormous volume of data generated by experiments to see this background. As such, a number of experiments are either beginning observations or completing construction. All of these mapping experiments are **interferometers**, in which the signals from multiple antennae are correlated with each other to produce one larger, higher-resolution telescope.

Currently, several experiments are either in the early phases of operations or final phases of construction. The wide ranges of approaches taken by the teams highlight the vitality of this field; the theoretical promise described in this chapter is now being transformed into actual instruments. The current tomographic projects include:

- The Giant Metrewave Radio Telescope (GMRT; in India) is an interferometer with thirty 45-m antennas operating at low radio frequencies. Completed over ten years ago, the 21-cm background was an early motivator for the project, but the theoretical landscape changed radically and only now has GMRT returned to this project. The large collecting area provides a powerful tool, but the instrument's narrow field of view and difficult radio environment present challenges. Nevertheless, the GMRT team was the first to put limits on the spin-flip background, ruling out a cold, neutral IGM at $z \sim 8$ in the summer of 2010.

- The Low Frequency Array (LOFAR; with the core in the Netherlands and outlying stations throughout Europe) is a large, general-purpose low-frequency radio telescope that began science operations in 2010. While its many other science goals mean that LOFAR is not completely optimized to observe the spin-flip background, its large collecting area (especially inside a compact "core" most useful for these observations) and powerful computers nevertheless make it a powerful machine for this purpose. Its location in Western Europe means that LOFAR will face by far the most difficult terrestrial radio environment. Moreover, it uses an enormous number of dipole antennae, combining their individual signals into "stations" that are then used as interferometers. While this allows for a large collecting area, it presents analysis challenges in understanding the instruments sufficiently well to extract the tiny cosmological signal.

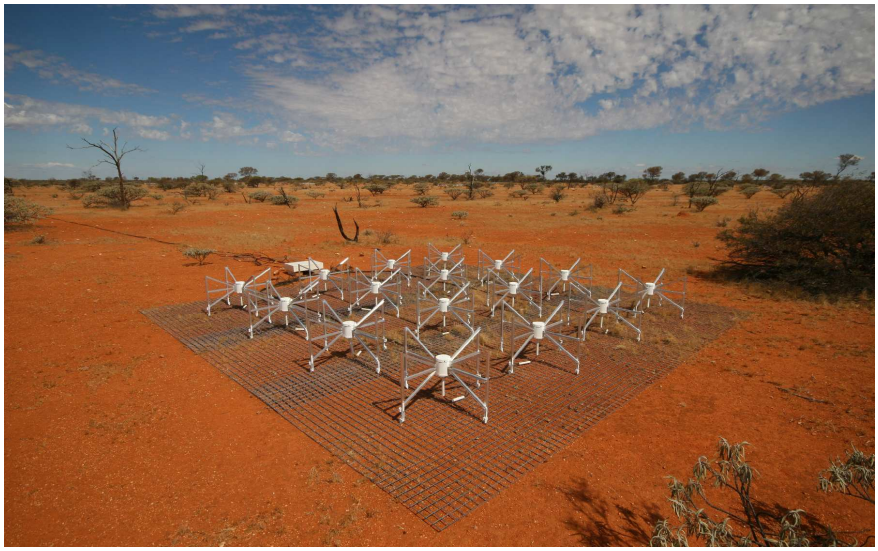

Figure 11.17 One of the antenna "tiles" used in the Murchison Widefield Array (MWA) experiment in Western Australia. Each such tile is composed of 16 crossed-dipole antennae, with their signals combined through hardware at the station. The full telescope combines the signals from the 500 tiles interferometrically. This allows for a large (several hundred square degree) field of view with a moderately large collecting area $\sim 8,000\ \mathrm{m}^2$ (comparable to the Very Large Array). The antennae operate between 80–300 MHz, corresponding to $z \approx 6$–15 (although the telescope will only be sensitive to the spin-flip background at $z < 12$). Image credit: C. Lonsdale.

- The Murchison Widefield Array (MWA in Western Australia) is an interferometer built almost entirely to observe the 21-cm background. As such, the project hopes to leverage the relatively small experiment into limits competitive with the largest first-generation experiments. Like LOFAR, MWA uses thousands of dipoles grouped into "tiles," which increase the collecting area at the cost of complexity. Because MWA's tiles are smaller, though, it achieves a larger field of view than LOFAR, which partially compensates for the much smaller collecting area. Figure 11.17 illustrates the antenna tile design of MWA.

- The Precision Array to Probe the Epoch of Reionization (PAPER, with instruments in Green Bank, West Virginia and South Africa) combines signals from single dipoles into an interferometer. Without tiles, PAPER, has a much smaller total collecting area but an advantage of a well-calibrated and well-understood instrument, along with an enormous field of view. The PAPER instrument is gradually building toward of order 100 antennas.

In addition to this impressive suite of ongoing efforts, larger experiments are planned for the future, with their designs and strategies informed by this present generation.

In this section we will briefly describe how these experiments work and hope to measure the spin-flip background. Of course we cannot hope to do full justice to a topic as rich as radio observations and interferometry in this chapter; we will focus on the ideas most relevant to the spin-flip background, and we refer the interested reader to one of the many good textbooks on radio astronomy for more detailed information.

11.7.1 A Brief Introduction to Radio Telescopes

The sensitivity of a telescope system depends on the competition between the strength of the cosmic signal collected by the antenna and the noise. Nevertheless, we will follow the usual conventions of radio astronomy in our notation. The signal output of the antenna can be specified as an **antenna temperature**, T_a, which is the temperature of a matched resistive load that would produce the same power level ($P_a = k_B\, T_a\, \Delta\nu$ for the resistor) as the signal power $P = A_e\, S_\nu\, \Delta\nu/2$ received in one of two orthogonal antenna polarizations, where S_ν is the source flux density (assuming an unpolarized source), $\Delta\nu$ is the observed frequency bandwidth, and A_e is the effective collecting area of the telescope. From these we define the antenna sensitivity factor $K_a \equiv T_a/S_\nu = A_e/2k_B$.

The signal-to-noise ratio is assessed by comparing T_a and $T_{\rm sys}$, the **system temperature**, similarly defined as the temperature of a matched resistor input to an ideal noise-free receiver that produces the same noise power level as measured at the output of the actual receiver. The system temperature includes contributions from both the telescope and receiver system and the sky; the latter dominates in our case. Noise fluctuations ΔT^N decline with increased bandwidth and integration time $t_{\rm int}$ according to the radiometer equation,

$$\Delta T^N = \kappa_c \frac{T_{\rm sys}}{\sqrt{\Delta\nu\, t_{\rm int}}} \approx \frac{T_{\rm sys}}{\sqrt{\Delta\nu\, t_{\rm int}}}, \tag{11.56}$$

where $\kappa_c \geq 1$ is an efficiency factor accounting for the details of the signal detection scheme; for simplicity we will set $\kappa_c = 1$, which is a reasonably approximation for the telescopes discussed here. The above equation has a simple interpretation. Since the occupation number of the photons is large, they behave as a classical electromagnetic wave. The number of independent samples of the noise temperature is then the number of cycles observed during the integration time $N_{\rm cyc} \sim \Delta\nu t_{\rm int}$, and the uncertainty in the system temperature $T_{\rm sys}$ is reduced by the factor of $\sqrt{N_{\rm cyc}}$ for Gaussian statistics (applicable in the limit of $N_{\rm cyc} \gg 1$).

The noise level (in flux density units) for an unresolved source is then

$$\sigma_S = \frac{T_{\rm sys}/K_a}{\sqrt{\Delta\nu\, t_{\rm int}}}. \tag{11.57}$$

Note that this decreases with the telescope collecting area A_e. However, in many applications, we must take into account that the total collecting area may be distributed over a much larger physical area, in order to achieve better angular resolution $\theta_D \approx \lambda/D_{\rm max}$, where λ is the (observed) wavelength and $D_{\rm max}$ is the

maximum distance between antennae. In this case, the equivalent brightness temperature uncertainty is

$$\Delta T^N = \frac{\sigma_S c^2}{2k_B \nu^2 \Omega_B} \equiv \frac{T_{\rm sys}}{\eta_f \sqrt{\Delta\nu\, t_{\rm int}}}, \tag{11.58}$$

where $\eta_f \equiv A_{\rm tot}/D^2_{\rm max}$ is the *array filling factor*. An appreciation of this dependence on η_f is crucial: the integration time required to detect a given surface brightness grows as $t_{\rm int} \propto D^4_{\rm max}$ if the (fixed) total collecting area is spread over larger areas in order to achieve better angular resolution.

We can develop better insight into the radio telescope response through a thought experiment in which a radio telescope is encased in a blackbody of temperature T. Regardless of its size, and with proper impedance matching, the telescope would produce an antenna temperature $T_a = T$ at its output. For this reason, attempts to observe the global 21 cm background are more concerned with issues of matching and gain calibration than with antenna size.

On the other hand, a telescope constructed with a beam of solid angle Ω_B will still deliver $T_a = T$ at its output if (i) it is embedded in a black body radiation field *or* (ii) an emitter of $T_B = T$ entirely fills its beam. Unfortunately, real radio telescopes do not form perfectly defined beams, and all suffer from sidelobes whose shapes and responses are dictated by diffraction and scattering of the incident radiation through the telescope. This is especially true of arrays, where a fraction $(1 - \eta_{\rm f})$ of the total response lies outside the beam defined by $\theta_{\rm D} \sim \lambda/D_{\rm max}$.

Using equation (11.55) with $T_{\rm sys} \approx T_{\rm sky}$ to estimate the telescope noise ΔT^N for a single-dish measurement of an unresolved source, we find

$$\Delta T^N|_{\rm sd} \approx 0.6\ {\rm mK} \left(\frac{1+z}{10}\right)^{2.6} \left(\frac{\rm MHz}{\Delta\nu}\,\frac{100\ {\rm hr}}{t_{\rm int}}\right)^{1/2}. \tag{11.59}$$

The mean 21-cm signal has $T_0 \sim 20\,{\rm mK}$; thus, single dish telescopes can easily reach the sensitivity necessary to detect the global 21-cm background.In this regime, the challenge is instead to separate the slowly varying cosmological signal from the foregrounds. Note, however, that detecting individual features is still limited by the resolution of the telescope: a small single dish can detect the mean signal across the entire sky but cannot identify individual ionized bubbles.

11.7.2 Noise Estimates for 21-cm Interferometers

At radio frequencies, interferometry is required to make maps with even a relatively coarse resolution; for realistic collecting areas, the array dilution factor η_f dramatically decreases the sensitivity. Again using equation (11.55) for the system temperature, we find

$$\Delta T^N|_{\rm int} \sim 2\ {\rm mK} \left(\frac{A_{\rm tot}}{10^5\ {\rm m}^2}\right) \left(\frac{10'}{\theta_D}\right)^2 \left(\frac{1+z}{10}\right)^{4.6} \left(\frac{\rm MHz}{\Delta\nu}\,\frac{100\,{\rm hr}}{t_{\rm int}}\right)^{1/2}. \tag{11.60}$$

The angular resolution scale of $\theta_D \sim 10'$ and the frequency resolution scale of $\Delta\nu \sim 1$ MHz correspond to ~ 20 comoving Mpc. More precisely, a bandwidth

$\Delta\nu$ corresponds to a comoving distance $\sim 1.8\,\mathrm{Mpc}(\Delta\nu/0.1\,\mathrm{MHz})[(1+z)/10]^{1/2}$, while an angular scale θ_D corresponds to $2.7(\theta_D/1')[(1+z)/10]^{0.2}$ Mpc. The current generation of telescopes have $A_{\rm tot} < 10^5$ m^2, so imaging will only be possible on large scales that exceed the typical sizes of bubbles during most of reionization. It is for this reason that near-term imaging experiments focus primarily on large H II regions generated by quasars in the middle phases of reionization, when the contrast between the large ionized bubble and the background IGM is largest.

Although equation (11.60) provides a simple estimate of an interferometer's sensitivity, we will see below that the rate at which interferometers sample different scales depends on its design; this effectively makes η_f a function of angular scale. Thus, equation (11.60) only provides a rough guide.

When two antennae are coupled together electronically to form an interferometer, the combined response projected on the sky resembles the characteristic diffraction pattern from a double slit. The spacing depends on the distance between the two elements, or the **baseline**. In general, the interferometer response to the sky brightness distribution $I_\nu(\hat{\mathbf{n}})$ for a particular "visibility" $\mathbf{V}$, corresponding to a particular baseline and frequency pair, in units of temperature, is

$$\mathbf{V}(\hat{\mathbf{n}}_0, u, v, \nu) \approx \int dx\, dy\, T_b(x, y, \nu) W_\nu(\hat{\mathbf{n}}_0, \hat{\mathbf{n}})\, e^{2\pi i\,(ux+vy)}, \qquad (11.61)$$

where W_ν is the normalized response pattern of the antennae and $\mathbf{A} = \lambda(u\hat{\mathbf{i}}, v\hat{\mathbf{j}}, w\hat{\mathbf{z}})$ is the vector (on the ground) between the two elements. In the orthogonal (u, v, w) coordinate system, the w axis aligns with the direction toward the sky at the center of the beam, and the u-v axes are oriented so that the v axis projects onto the local meridian. The coordinates x and y are angles measured in the "sky plane" relative to the intersection of $\hat{\mathbf{z}}$ with the celestial sphere. In this Fourier transform of the sky, u and v represent spatial frequencies and the w axis produces a phase offset in the interferometer fringe that can be calibrated. (This representation assumes that the interferometer sees only a small piece of the sky so that the "flat sky" approximation is valid; that is not actually true for some of the 21-cm telescopes, but the basic formalism presented here provides a reasonable approximation with much less technical difficulty.)

We must keep in mind that this Fourier integral does not properly account for sources far outside the primary beam; in effect, these add a noise-like contribution entering through the sidelobes that inevitably appear outside the primary beam.

Given the difficulty of high signal-to-noise imaging, attention has focused on statistical measurements. We will now turn to estimating the sensitivity of 21 cm experiments to the power spectrum. Error estimates for other statistical measures must still be developed, but the basic principles are the same. For simplicity, we will only consider the effects of thermal noise and cosmic variance, which provide a fundamental limit. Systematics (especially foregrounds) present equally large difficulties, and the community is hard at work developing strategies to mitigate them, some of which we will discuss below.

We begin with the complex visibility of equation (11.61). The detector noise for a single visibility measurement is closely related to equation (11.56). In the regime

of $T_{\rm sys} \approx T_{\rm sky}$, equation (11.57) implies

$$\Delta T^N(\nu) = \frac{\lambda^2\, T_{\rm sky}}{A_e\, \sqrt{\Delta\nu t_{\mathbf{u}}}}, \tag{11.62}$$

where here $t_{\mathbf{u}}$ is the integration time of this particular baseline; due to the Earth's rotation these large interferometers continually shift their coverage (in an analogous manner to "drift-scanning" in optical astronomy) so this is *not* the same as the total integration time. Also, A_e is the collecting area of each antenna element (which we assume to be perfectly efficient, for simplicity).

The observed "visibility data cube" is actually a hybrid of Fourier space ($u,\, v$) and redshift-space (ν) coordinates and is thus inconvenient for comparing to theoretical models. One can either transform the visibility data to the sky plane to obtain the "image cube" or transform the frequency (redshift) coordinate to its Fourier-space equivalent in order to obtain a representation with spatial frequency for all three dimensions,

$$T_b(\mathbf{u}) = \int_B d\nu\, \mathbf{V}(u, v, \nu)\, e^{2\pi i \eta\nu}, \tag{11.63}$$

where the integration extends over the full bandwidth B of the observation, $\mathbf{u} \equiv u\hat{\mathbf{i}} + v\hat{\mathbf{j}} + \eta\hat{\mathbf{z}}$, and η has dimensions of time. In this representation, the effective noise can be obtained by Fourier transforming the signal across the frequency axis, yielding

$$\Delta T^N(\mathbf{u}) = \frac{\lambda^2\, T_{\rm sys}\, \sqrt{B}}{A_e\, \sqrt{t_{\mathbf{u}}}} \approx \frac{T_{\rm sys}}{\sqrt{B\, t_{\mathbf{u}}}} \times \frac{\lambda^2}{A_e\, \delta\eta}. \tag{11.64}$$

In the second equality, we have set $\delta\eta$ equal to the inverse bandwidth. The factor $A_e/\lambda^2 \times \delta\eta$ then represents the Fourier space resolution of the observation (or the inverse volume sampled by the primary beam, in the appropriate units); note the similarity to equation (11.56) when written in this form. Here $\Delta T^N(\mathbf{u})$ has units of temperature divided by time, because of the Fourier transform in the frequency direction.

To estimate the statistical errors, we need the covariance matrix of the noise for antenna pairs at baselines $\mathbf{u}_i$ and $\mathbf{u}_j$. Because the thermal noise errors are uncorrelated between measurements, this is simply a diagonal matrix with each element being the square of equation (11.64). In transforming to the physical wavevector $\mathbf{k}$, we distinguish between the component $\mathbf{u}_\perp$ oriented along the sky (corresponding to $\mathbf{k}_\perp = 2\pi\mathbf{u}_\perp/D$, where D is the comoving distance to the observed survey volume) and the component $\mathbf{k}_\parallel$ along the line of sight. This is useful because interferometers can have arbitrarily good frequency resolution while the $\mathbf{u}_\perp$ coverage is fixed by the baseline distribution.

We define the number density of baselines that observe a given $\mathbf{u}_\perp$ as $n(\mathbf{u}_\perp)$; this is normalized so that its integral over the half-plane is $N_B = N_a\,(N_a - 1)/2$, the total number of baselines in the array of N_a antennae. Two properties of $n(\mathbf{u}_\perp)$ are noteworthy. First, because of the earth's rotation, it is azimuthally symmetric and only a function of $u_\perp = |\mathbf{u}_\perp|$. Second, for a smooth antenna distribution, $n(u_\perp)$ is virtually always a decreasing function of $u_\perp$. This follows from a simple

geometric consideration: it is difficult to arrange the antenna distribution to have many more long baselines than short ones. We can write:

$$t_{\mathbf{k}} \approx n(u_\perp) \left(\frac{A_e}{\lambda^2}\right) t_{\rm int}. \tag{11.65}$$

As before $A_e/\lambda^2 \approx \delta u\, \delta v$ is the angular component of the Fourier-space resolution. Thus, the noise covariance matrix is

$$\begin{aligned} C^N(\mathbf{k}_i,\, \mathbf{k}_j) &\equiv \left\langle \Delta T^N(\mathbf{u}_i)^*\ \Delta T^N(\mathbf{u}_j) \right\rangle \\ &= \left(\frac{\lambda^2\, B\, T_{\rm sys}}{A_e}\right)^2 \frac{\delta_{ij}}{B\, t_{\mathbf{k}}}. \end{aligned} \tag{11.66}$$

Equation (11.66) represents the thermal noise contribution to the covariance matrix; even in an ideal experiment with no systematics from foregrounds, we must also include errors from sample variance. This component is

$$\begin{aligned} C^{SV}(\mathbf{k}_i,\, \mathbf{k}_j) &= \langle T_b^*(\mathbf{k}_i)\, T_b(\mathbf{k}_j) \rangle \\ &\approx \delta_{ij} T_0^2 \left\langle x_H \right\rangle^2 \int d^3\mathbf{u}\, |\tilde{W}(\mathbf{u}_i - \mathbf{u})|^2\, P_{21}(\mathbf{u}) \\ &\approx T_0^2 \left\langle x_H \right\rangle^2 P_{21}(\mathbf{k}_i)\, \frac{\lambda^2\, B^2}{A_e\, D^2\, \Delta D}\, \delta_{ij}, \end{aligned} \tag{11.67}$$

where $\Delta D \propto B$ is the line-of-sight depth of the observed volume in comoving units. In the first line, the average is over baseline and frequency pairs indexed by $\mathbf{k}_i$ and $\mathbf{k}_j$ (or equivalently $\mathbf{u}_i$ and $\mathbf{u}_j$). In the second line, $\tilde{W}$ is the Fourier-transform of the primary beam response function, including the finite bandwidth, and is most naturally expressed in the "observed" units $\mathbf{u}$. It typically differs from zero in an area $\delta u\, \delta v\, \delta\eta \approx A_e/(\lambda^2 B)$ and (ignoring efficiencies) integrates to unity over the beam. For the last line, we have assumed that $\mathbf{u}$ is much larger than the width of this response function. Then $P_{21}(\mathbf{u})$ is constant across the beam and can be pulled out of the integral, which becomes simply $(\delta u\, \delta v\, \delta\eta)^{-1}$. We have also transformed to the more physically relevant wavenumber $\mathbf{k}$, which introduces a factor $B/(D^2 \Delta D)$.

Equation (11.67) has a simple physical interpretation: it is essentially a normalization factor $(T_0^2 \left\langle x_H \right\rangle^2 B^2)$ multiplied by $P_{21}/V_{\rm surv}$, where $V_{\rm surv} \approx D^2 \Delta D (\lambda^2/A_e)$ is the total volume observed by the telescope. The second factor counts the number of independent estimates N available to the measurement of a given Fourier mode; the squared error then scales as $1/N$.

To translate these into error estimates, we use the common Fisher information matrix approach. This provides an idealized estimate of the measurement errors given the total covariance matrix $\mathbf{C} = \mathbf{C}^N + \mathbf{C}^{SV}$, ignoring any possible systematics and inefficiencies in the data reduction. Given a vector of parameters $\mathbf{\Psi}$, the $(i,\, j)$ element of the Fisher matrix is defined as

$$F_{ij} \equiv \left\langle -\frac{\partial^2 \ln \mathcal{L}}{\partial\Psi_i\, \partial\Psi_j} \right\rangle \tag{11.68}$$

$$= {\rm Tr}\left[\mathbf{C}^{-1}\, \frac{\partial \mathbf{C}}{\partial \Psi_i}\, \mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \Psi_j}\right], \tag{11.69}$$

where $\mathcal{L}$ is the log-likelihood function. For the simple case of measuring the binned power spectrum from the data points, the "parameters" are the power spectrum amplitudes in each of the bins, $\Psi_i = P_{\Delta T} \equiv T_0^2 \langle x_H \rangle^2 P_{21}(\mathbf{k}_i)$; in more general cases they are the parameters of a theoretical model meant to describe the data. The Cramer-Rao inequality states that the errors on any unbiased estimator of the power spectrum must satisfy

$$\delta P_{\Delta T}(\mathbf{k}_i) \geq \frac{1}{\sqrt{N_c(\mathbf{k}_i)}} \sqrt{(\mathbf{F}^{-1})_{ii}}, \tag{11.70}$$

where N_c is the number of measurements in the appropriate bin and $\mathbf{F}^{-1}$ is the inverse of the Fisher matrix.

In the case we are studying, the Fisher matrix is particularly simple to use because the covariance matrix is diagonal. (This will not be true for real data, because foreground cleaning and other systematic effects induce correlated residual errors, but it provides a rough estimate of the noise limits.) The resulting error (from a single baseline) on a power spectrum estimate is

$$\delta P_{21}(\mathbf{k}_i) = P_{21}(\mathbf{k}_i) + \frac{T_{\rm sys}^2}{B t_{\rm int}} \frac{D^2 \Delta D}{n(k_\perp)} \left(\frac{\lambda^2}{A_e}\right)^2. \tag{11.71}$$

The last step is to count the number of Fourier cells in each bin, which depends on the Fourier-space resolution of the instrument. Recall that when redshift space distortions are included P_{21} is not truly isotropic, but it is azimuthally symmetric. Thus, we use Fourier cells grouped into annuli of constant $(k,\, \mu)$. Then, in the limit that $k\mu$ is much larger than the effective $k_\parallel$ resolution,

$$N_c(k) \approx 2\pi\, k^2\, \Delta k\, \Delta\mu \times \left[\frac{V_{\rm surv}}{(2\pi)^3}\right], \tag{11.72}$$

where the last term represents the Fourier space resolution. The total errors from all estimates within a bin simply add in quadrature.

Equations (11.66), (11.67), and (11.71) fully specify the effects of noise in the absence of systematic effects. But to make estimates we must determine the effective observing time $t_{\mathbf{k}}$ for each mode – and hence the baseline distribution $n(u_\perp)$ by equation (11.65) – as well as the sampling density (Eq. 11.72 for a measurement in annuli). These two quantities are obviously highly dependent on the design of the experiment. It is therefore useful to consider the simple thermal noise-dominated case in order to develop some intuition for array design. Substituting for N_c in equation (11.71) and ignoring the first term, we find

$$\delta P_{\Delta T} \propto A_e^{-3/2}\, B^{-1/2} \left[\frac{1}{k^{3/2}\, n(k,\, \mu)}\right] \left(\frac{T_{\rm sys}^2}{t_{\rm int}}\right). \tag{11.73}$$

Here we have assumed that the power spectrum is measured in bins with constant logarithmic width in k but constant linear width in μ. From equation (11.73), we can deduce a number of fundamental considerations driving array design.

- First, $\delta P_{21} \propto t_{\rm int}^{-1}$, because the power spectrum depends on the square of the intensity.

- Second, we can increase the collecting area in two ways. One is to add antennae while holding the dish area A_e constant. Recall that $n(k, \mu)$ is normalized to the total number of baselines $N_B \propto N_a^2$: thus, adding antennae of a fixed size decreases the errors by the total collecting area squared. (Of course, the number of correlations needed also increases by the same factor, so this strategy has other costs.) The other method is to make each antenna larger but hold their total number fixed. In this case, the total number of baselines, and hence $n(k, \mu)$, remains constant, but $\delta P_{\Delta T} \propto A_e^{-3/2}$. Increasing the collecting area in this way is not as efficient because it decreases the total field of view of the instrument.

- Third, adding bandwidth increases the sensitivity relatively slowly: $\delta P_{\Delta T} \propto B^{1/2}$, because it adds new volume along the line of sight without affecting the noise on any given measurement. Of course, one must be wary of adding too much bandwidth because of systematics (especially foregrounds).

- Finally, as a function of scale k, $\delta P_{\Delta T} \propto k^{-3/2}\, n(k, \mu)^{-1}$. The first factor comes from the increasing (logarithmic) volume of each annulus as k increases. But in realistic circumstances the sensitivity actually decreases toward smaller scales because of n. This is most obvious if we consider a map at a single frequency. In that case, high-k modes correspond to small angular separations or large baselines; for a fixed collecting area the array must therefore be more dilute and the sensitivity per pixel decreases as in equation (11.60). In the (simple but unrealistic) case of uniform uv coverage, the error on a measurement of the angular power spectrum increases like θ_D^{-2} for a fixed collecting area.

  Fortunately, the three-dimensional nature of the true 21 cm signal moderates this rapid decline toward smaller scales: even a single dish can measure structure along the line of sight on small physical scales. Mathematically, because $n(k, \mu) = n(k_\perp)$, each baseline can image arbitrarily large $k_\parallel$, at least in principle. For an interferometer, this implies that short baselines still contribute to measuring large-k modes. Thus, provided that they have good frequency resolution, compact arrays are surprisingly effective at measuring small-scale power . There is one important caveat: if short wavelength modes are only sampled along the frequency axis, we can only measure modes with $\mu^2 \approx 1$. Thus we recover little, if any, information on the μ dependence of the redshift-space distortions. Studying this aspect of the signal *does* require baselines able to measure the short transverse modes with $\mu^2 \approx 0$.

Figure 11.18 summarizes the expected errors (including only thermal noise and cosmic variance, not systematics) on the spherically-averaged power spectrum. The thin curves show a forecast for an experiment like the MWA (though all of the first-generation experiments have similar statistical power). We assume 1000 hours of integration on a single field (roughly one-year of realistic observing conditions), bin the observed modes into segments of width $\Delta k = k/2$ (as shown by the horizontal bars), and take a radial survey width corresponding to 6 MHz (or $\Delta z \sim 0.5$).

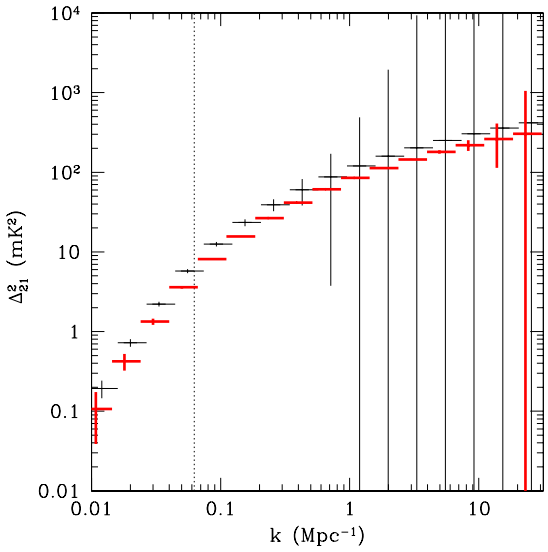

Figure 11.18 Estimated errors on the spherically-averaged 21-cm power spectrum at $z = 8$ for an MWA-like experiment (thin curves) and one with 100 times larger collecting area (thin curves). The central values of the latter are shifted downward for clarity of presentation. We assume 1000 hours of integration on a single field, bin the observed modes into segments of width $\Delta k = k/2$ (the horizontal bars show these bins), and take a radial survey width corresponding to 6 MHz (or $\Delta z \sim 0.5$). The vertical dotted line shows the scale corresponding to this bandwidth; foreground removal will likely prevent measurements at wavenumbers smaller than this effective width.

Provided that it reaches this limit, the MWA can place fairly stringent constraints at scales $k < 1$ Mpc^{-1}. Smaller scales are swamped by thermal noise. The errors on large scales come from cosmic variance, although here they are quite modest because of the large field of view of the telescope. Reaching smaller scales will require more collecting area in order to reduce the noise. The thick curves show the estimated errors for a futuristic experiment, with 5000 antenna tiles (ten times more than the MWA) and one hundred times larger total collecting area. This experiment would provide good constraints out to $k \sim 10$ Mpc^{-1}.

Unfortunately, measuring very large physical scales with these experiments is likely to be very difficult, because of the other astronomical foregrounds. To separate the Galactic synchrotron radiation from the cosmological signal, the experiments will rely on the former's spectral smoothness and the latter's rapid variations with frequency (due to H II regions, density fluctuations, or temperature variations). The essential idea is to fit a low-order function to each pixel in the map (or Fourier mode) and subtract out this mean variation over a wide (several MHz) frequency range. Provided that the foregrounds are smoother than the signal, this scheme will isolate the spin-flip background, but with an inevitable loss of infor-

mation (i.e., any variations in the 21 cm background over large frequency ranges are also subtracted out). Current estimates suggest that this method will work very well at small scales but prevent measurements of any fluctuations on scales larger than those corresponding to the several MHz bandwidth of each measurement. The vertical dotted line in Figure 11.18 shows the scale corresponding to our assumed 6 MHz bandwidth; modes to the left of this line are likely lost in the foreground removal process. Unfortunately, this drastically reduces the dynamic range of the first generation experiments.

Because the sky noise increases rapidly with redshift (see eq. 11.55), the first generation of experiments lose sensitivity at $z \sim 11$–12. Reaching these high redshifts will likely require collecting areas approaching a square kilometer. Such large instruments will also be necessary to measure the redshift-space distortions in the spin-flip background, because they require separate measurements of power along the line of sight and across the plane of the sky. The first generation experiments are relatively small and do not have adequate sensitivity to make high-resolution measurements on the plane of the sky, although they can do so in the redshift direction with reasonably narrow frequency channels. Much larger instruments are necessary to build sensitivity to fluctuations on the plane of the sky.

In addition to the unavoidable problems posed by foreground cleaning, there are several other serious systematic challenges before one can reach the limits suggested by Figure 11.18. These include the ionospheric refraction described earlier, the many bright astronomical point sources (and especially their sidelobe contamination to the antenna beam), the variation of the instrument properties with frequency, and the polarized component of the foregrounds (which can vary rapidly with frequency and hence escape the typical foreground removal algorithms). Fortunately, the several experiment teams are each tackling these problems in unique ways, and the community hopes that they can be overcome in the near future.

Chapter Twelve

Other Probes of the First Galaxies

So far we have discussed three classes of observational probes of the first galaxies: direct observations of individual galaxies (over a variety of wavelengths), the Lyman-α line (both as a test of the galaxy populations and the IGM), and the spin-flip line from intergalactic gas. However, there are many other, less direct, ways to probe structures during the cosmic dawn. In this chapter, we will discuss several of these, including

- *Secondary anisotropies* of the cosmic microwave background (CMB) as those photons pass through gas during the cosmic dawn. CMB photons do not interact with the IGM gas until it is ionized; however, once that occurs, the photons begin to scatter off the free electrons. The scattering process induces both large-scale polarization and small-scale temperature anisotropies.

- *Diffuse backgrounds* from the cosmic dawn (other than the spin-flip background) can result from the integrated emission of the entire galaxy population. Typically, these backgrounds include galactic emission lines, ranging from CO lines in the radio to the Lyman-α line itself, so (like the spin-flip background) they contain not only angular structure but also spectral structure. Measuring these integrated backgrounds via low-resolution observations can be much easier than detecting individual galaxies (though of course also contains less information) and can reveal useful information about the global galaxy populations.

  The *cross-correlation* of different probes can help to isolate cosmological information in the presence of contaminants and can often isolate interesting aspects of the signal.

- *Gravitational waves* from black hole mergers in the early Universe are potentially detectable with future interferometers. This new observational window can inform us about the hierarchical growth of the first generation of supermassive black holes.

- *Fossil structure* from early galaxies that remains in (or can be deduced from) the Milky Way or other nearby structures in the Local Group. This includes the residual effects of feedback (from the Lyman-Werner background, photoheating, or other processes) on the small satellite galaxies or globular clusters of the Milky Way, old low-mass stars that may have formed during the cosmic dawn and remain the Milky Way (or its halo), and remnant signatures of the early merger history of the Milky Way.

12.1 SECONDARY COSMIC MICROWAVE BACKGROUND ANISOTROPIES FROM THE COSMIC DAWN

The CMB indicates that hydrogen atoms formed 400,000 years after the Big Bang, as soon as cosmological expansion cooled the gas below 3,000 K. Once neutral, the CMB photons could only interact with the IGM gas through its resonant transitions – and after $z \sim 1000$, when the photons redshift out of the Lyman-α resonance, the only such available transition is the 21-cm line, whose effects we have already examined.

However, once the first stars or black holes begin to ionize the IGM, the CMB photons began to scatter off the free electrons. This has several effects on the CMB anisotropies, which we will describe in detail here.

12.1.1 Large-Scale Polarization of the CMB

The crucial parameter in determining the effects is the total CMB optical depth to electron scattering,

$$\tau_{\rm es} = \int n_e(z)\sigma_T(cdt/dz)dz, \tag{12.1}$$

where $\sigma_T = 6.65 \times 10^{-25}\,{\rm cm}^2$ is the Thomson scattering cross section and the integral is over the path taken by a photon; note that only redshifts where the ionized fraction is non-zero will contribute (in fact, the residual ionized fraction following recombination produces only a very small contribution to $\tau_{\rm es}$, so the integrand is only significant once reionization begins). In the simplest approximation, where we assume that he IGM is instantaneously ionized at $z_{\rm reion}$, equation (12.1) can be integrated analytically, yielding

$$\tau_{\rm es} = 4.75 \times 10^{-3} \times \{[\Omega_\Lambda + \Omega_m(1+z_{\rm reion})^3]^{1/2} - 1\} \tag{12.2}$$

The most obvious effect of this scattering on the CMB temperature anisotropies is to wash them out, as a fraction $e^{-\tau}$ of the photons that appear to be incident from a particular direction actually come from elsewhere. However, each line of sight samples only photons from a finite surrounding region, the causal horizon. Thus, the angular power spectrum of the CMB fluctuations, C_ℓ – which contains two factors of temperature – is damped by a factor $e^{-2\tau}$ on angular scales smaller than the horizon at that time. On larger scales, the power spectrum is unaffected. Unfortunately, this slight change to the slope of the temperature anisotropy power spectrum is not observable.

However, the scattering process also induces polarization, which is observable. Consider photons that scatter off an electron in the $+z$ direction (where the observer sits), as shown in Figure **??**. Photons incident from the $\pm x$ directions can scatter in the $+z$ direction only if they are polarized along the plane of the page in the horizontal direction, because the other possible polarization state of the incident wave – in the xz plane, out of the page – would not produce a scattered transverse wave. In contrast, to produce a transverse wave photons incident from the $\pm y$ directions can scatter toward the observer only if they are polarized vertically in the figure (along the plane of the page).

If we suppose that the electron scatters photons from all directions, it will produce a mixture of these horizontal and vertical polarization states. However, if there is an asymmetry in the incident radiation field – in particular, a quadrupole anisotropy between the intensity of the background along the $\pm x$ and $\pm y$ axes – the resulting mixture will have a net polarization.

For the CMB, the electrons during the cosmic dawn can scatter any photons incident upon them that originated within the causal horizon at that time. The net polarization is $P \sim 0.1\tau_{\rm es}Q$, where Q is the primordial quadrupole anisotropy on that horizon scale (the prefactor 0.1 comes from the detailed physics of the Thomson scattering process). The overall amplitude of the polarization (or its power spectrum) therefore provides a measure of $\tau_{\rm es}$ and hence the integrated column of ionized gas between us and the recombination surface.

The first such measurements have been made with the *Wilkinson Microwave Anisotropy Probe (WMAP)* over the past several years. The current estimate is $\tau_{\rm es} = 0.087 \pm 0.017$, which would correspond to instantaneous reionization at $z \sim 10$. Unfortunately, $\tau_{\rm es}$ on its own does little to distinguish different reionization histories, as it simply measures the total amount of scattering between the present day and the surface of last scattering. This number will be refined by forthcoming CMB data from the Planck satellite.[i]

To learn about the history, we must turn to the scale dependence of the polarization, or equivalently the shape of its power spectrum. The polarization anisotropies then appear to us on the angular scale subtended by this horizon distance, which occurs at $\ell < 40$. On much finer angular scales, the post-reionization scattering actually washes out any primordial polarization, just as it does to the temperature anisotropies. However, because there is no primordial polarization on large scales anyway, the "reionization bump" from the late-time scattering is very clearly distinguishable from primordial anisotropies.

Interestingly, because this horizon scale evolves with redshift, there is some information about the time history of $\bar{x}_i(z)$ contained in the CMB polarization power spectrum. Figure 12.1 shows some example models (left) and their observable signatures (right). The models were generated with simple prescriptions for the physics of the first galaxies, but their origins are not important for our purposes; they can be taken simply as contrasting reionization histories to gauge its effects on the shape and amplitude of the polarization power.[ii] Models 1–3 hold $\tau_{\rm es} = 0.17$ but vary $\bar{x}_i(z)$. Models 4 and 5 show larger and smaller overall optical depths (of 0.23 and 0.14, respectively). Note that all of these values are considerably larger than the current best estimate from *WMAP* but are only shown here to illustrate the dependence of the polarization signal on the reionization history.

The right panel shows the corresponding power spectra of the polarization anisotropies (in detail it shows the EE, or scalar, component, which contains most of the contributions from reionization). Models 1–3 (shown by the dark lines) have very amplitudes but different detailed shapes, particularly around the trough at $\ell \sim 20$–

[i] http://www.rssd.esa.int/index.php?project=planck

[ii] Model 3 in particular is physically implausible, because a non-monotonic $\bar{x}_i(z)$ requires very strong, instantaneous feedback suppressing galaxy formation.

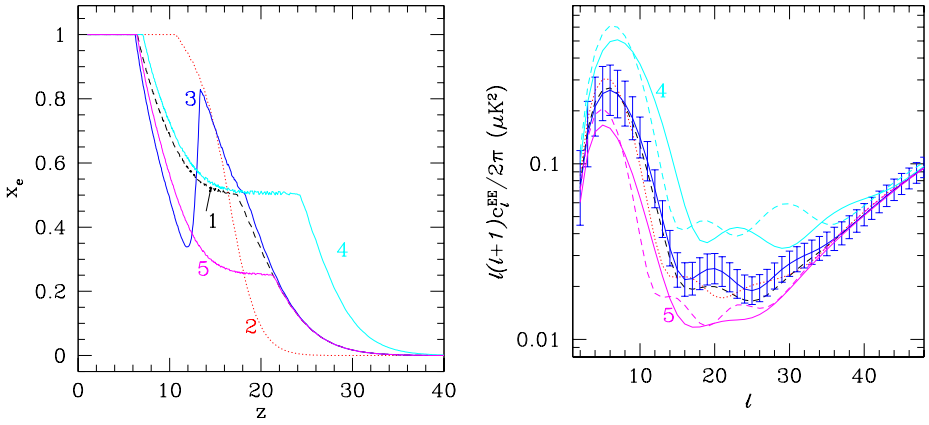

Figure 12.1 *Left:* Five example models of the reionization history. Models 1–3 are normalized to have $\tau_{\rm es} = 0.17$, while models 4 and 5 have $\tau_{\rm es} = 0.14$ and 0.23, respectively. Note that all these values are well above the current best estimate. *Right:* CMB polarization power spectra for these five models (shown is the EE component, which includes scalar perturbations). All are normalized to have the same power at $\ell > 50$, where reionization has no significant effect. The bold-dashed, dashed, and solid lines correspond to models 1–3; the latter has cosmic variance error bars attached.The light solid lines show models 4–5, while the light dashed lines show best fit polarization spectra for instantaneous reionization models. Figure credit: Holder & Haiman 2003, ApJ, 595, 13.

30; the error bars, which show the ideal cosmic variance errors, show that these models can at least in principle be separated at high confidence. It is much easier to separate models with different optical depths: models 4 and 5 are shown by the light solid lines. Here we can also see the effect of $\bar{x}_i(z)$ on the power spectrum: the light dashed lines show the best fit instantaneous reionization models, which provide relatively poor fits to the more complex reionization histories used in these models.

The reionization era also generates anisotropies in the B-mode polarization (so-called because they have non-zero curl, as opposed to the E-modes, which are curl-free). These type of anisotropies are of particular astrophysical interest because they are also generated by the gravitational wave background from the inflationary era, and there are numerous efforts underway to measure their properties. Secondary fluctuations, like those generated by reionization, are therefore important contaminants to understand. Fortunately, it appears that the B-modes generated by reionization are small and have large angular coherence in the polarization direction, which make them relatively easy to isolate. Moreover, the 21-cm background can be used to reconstruct the electron-scattering optical depth along different lines of sight and from that to reconstruct the expected polarization pattern (see §12.3). Thus, B-mode anisotropies from reionization should not pose a substantial chal-

lenge to the detection of primordial gravitational waves.

The *Planck* satellite should be able to distinguish different reionization scenarios using the polarization anisotropies over the next several years. However, this technique is only sensitive to the global reionization history, not the details of the reionization process. To understand the growth and morphology of the ionized regions, we must probe much smaller angular scales.

12.1.2 Secondary Temperature Anisotropies

On small scales, inhomogeneities in the density, ionized fraction, and velocity field combine to produce temperature anisotropies during reionization. These anisotropies are referred to as the *kinetic Sunyaev-Zel'dovich (kSZ) effect*, which broadly encompasses two distinct physical components, the Ostriker-Vishniac effect and the patchy reionization signal.

Let us first consider the general expression for the temperature along a line of sight $\hat{\mathbf{n}}$,

$$\frac{\Delta T(\hat{\mathbf{n}})}{T} = \int d\eta\, e^{-\tau_{\rm es}(\eta)} a n_e(\eta) \sigma_T \frac{\hat{\mathbf{n}} \cdot \mathbf{v}(\eta)}{c}, \tag{12.3}$$

where $\eta = \int_0^t dt'/a(t')$ is the conformal time, $\mathbf{v}$ is the peculiar velocity of the ionized gas and $\tau_{\rm es}(\eta)$ is the optical depth between the observer and a conformal time η. (The extra factor a occurs because $cdt = ad\eta$.) Perturbations in the temperature will then be sourced by the product of the peculiar velocity along the line of sight and the ionized gas density; we define $\mathbf{q} = \mathbf{v}(1 + \delta_b + \delta_x)$ for convenience. Here we must be careful to include the baryonic density fluctuation (rather than cold dark matter) because it is this material that scatters the photons.

This equation indicates that only the component of $\mathbf{v}$ along the line of sight contributes to the anisotropy. Suppose then that we work to linear order, so that $\mathbf{q} \approx \mathbf{v}$ (since the peculiar velocity is itself a first order quantity). Because the Fourier transform of this velocity is parallel to the wavevector $\mathbf{k}$ (see eq. 2.10), this implies that only those modes along the line of sight contribute. But for such waves oriented along the line of sight, their troughs and crests will (nearly) cancel, especially on small angular scales where there are many such troughs and crests.

Thus, angular correlations can only be generated by modes perpendicular to the line of sight – which do not in the end produce any correlation from the linear order velocity term – and from components of $\mathbf{q}$ perpendicular to the wavevector, which we will call $\mathbf{q}_\perp$. But the latter cannot be generated by the velocity itself, so it is the *nonlinear* terms $\delta_b \mathbf{v}$ and $\delta_x \mathbf{v}$ that generate anisotropies. Because the CMB is projected onto the sky, the angular fluctuation power spectrum is therefore

$$C_\ell = (\sigma_T \bar{n}_e^c)^2 \int \frac{d\eta}{r^2} W(\eta)^2 P_{q\perp}(\ell/r, \eta), \tag{12.4}$$

where $\bar{n}_e^c$ is the comoving free electron density (assuming full ionization), r is the comoving distance from the observer to a conformal time η, $W = \bar{x}_i e^{-\tau_{\rm es}}/a^2$, and $P_{q\perp}$ is the three-dimensional power spectrum of the projection $\mathbf{q}_\perp$. This integral picks out the physical scales corresponding to a given observed multipole moment,

weighting the contribution from each redshift by the factor W^2/r^2, which includes both the IGM ionized fraction (or the fraction of matter that can actually scatter CMB photons at the relevant redshift) and the effect of subsequent scattering washing out the secondary anisotropies (via the exponential).

The power spectrum $P_{q\perp}$ involves four-point functions (or correlations between four quantities) in many possible combinations, some of which are negligible. For example, in practice these four-point functions factor into pairs of normal power spectra, because the "connected" higher-order correlations vanish for Gaussian random fields. Moreover, terms like $P_{\delta v}$ can be ignored because of the scale mismatch between these two quantities (recall that, in linear theory, $v \propto \delta/k$ so is driven by large scale modes – while the δ fluctuations of interest occur only on small scales).

In most models, the dominant contribution to $P_{q\perp}$ is

$$P_{\rm OV} = \frac{1}{3} v_{\rm rms}^2 P_{\delta_b \delta_b}, \tag{12.5}$$

which describes the *Ostriker-Vishniac effect*[42], which arises from scattering off ionized clouds with bulk motions. The rms velocity is given by

$$v_{\rm rms}^2 = \int dk \, \frac{k^2}{2\pi^2} P_{vv}(k). \tag{12.6}$$

Because this effect is most important on small scales, one must usually use the nonlinear density power spectrum (filtered appropriately for small-scale baryonic structure) to evaluate the $P_{\delta_b \delta_b}$ contribution – either through numerical simulations or an approximation like the halo model.

The Ostriker-Vishniac effect has contributions from all redshifts at which the IGM (or even gas near or within galaxies) is ionized, and it is strongest at lower redshifts, where both the rms velocity field and the density fluctuations are most significant. Its total amplitude does, however, depend on when reionization began, and its shape depends very slightly on this as well (both because of the changing angular diameter distance and the evolving characteristic scale of structure formation).

The left panel of Figure 12.2 shows some example angular power spectra of the Ostriker-Vishniac effect in three different models of reionization; in each case we take $Q_{\rm HII} \propto f_{\rm coll}$ and calibrate so that reionization ends at $z_{\rm reion} = 8$, 12, and 18 (thick dashed, solid, and dot-dashed curves, respectively). Note that, even with these rather different histories, the Ostriker-Vishniac signal only changes by $\sim 10\%$. Thus, deducing information about reionization from this component of the CMB anisotropies will be challenging, requiring very careful modeling of the dominant lower-redshift contributions.

The other two terms that contribute to $P_{q\perp}$ involve integrals over $P_{vv} P_{\delta_x \delta_x}$ and $P_{vv} P_{\delta \delta_x}$, which are known as the "patchy reionization" contributions. Because these involve fluctuations in the ionized fraction, they are only relevant during the reionization era itself and so better isolate that period's properties. Physically, they originate from the peculiar velocities of the ionized bubbles that appear during reionization, which have biased velocities relative to the background. They can actually be estimated analytically using the simple tools describing the statistical

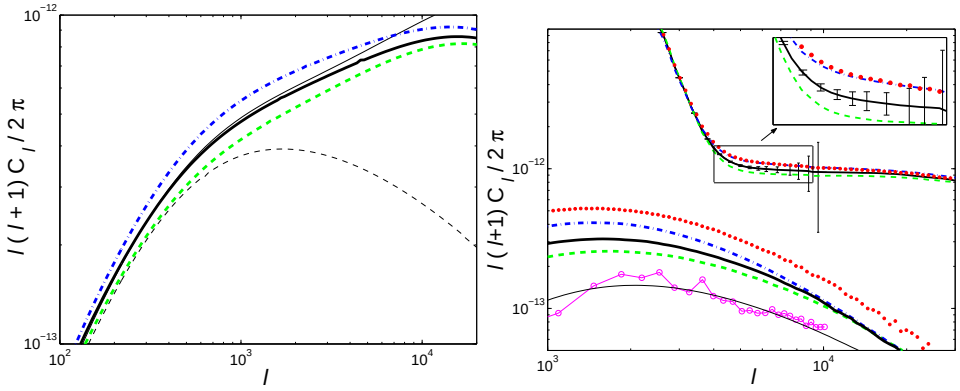

Figure 12.2 *Left:* Angular power of the Ostriker-Vishniac effect on the CMB. The thick dashed, solid, and dot-dashed curves take models in which reionization ends at $z_{\rm reion} = 8$, 12, and 18, respectively. The thin solid curve shows the signal without any baryonic filtering, while the thin dashed curve shows the signal assuming only a linear theory density field. *Right:* The patchy reionization signal from the same models (lower left, thick curves) and the total CMB anisotropy power spectrum (at upper right, including the primordial, lensing, and Ostriker-Vishniac components as well). In addition, the dotted curve shows the patchy signal from a model with extended reionization. The open circles show the estimate from a semi-numeric simulation of reionization, with the thin line a corresponding analytic estimate with the same reionization history; the two match very well. The error bars at upper right are representative of current instruments, assuming perfect foreground removal. Figure credit: McQuinn et al. 2005, ApJ, 630, 643.

properties of the ionization field developed from our excursion set model of reionization in §8.8.

The patchy reionization contribution therefore depends both on the structure of the ionized bubbles and their relation to the density field; the right panel of Figure 12.2 shows some example models. The thick dashed, solid, and dot-dashed curves at bottom left show this component of the signal for the same models as in the left panel. In all cases, the signal peaks at $\ell \sim 2000$, which is simply the angular scale corresponding to the projected physical sizes of the bubbles midway through reionization (where the patchiness peaks). The different amplitudes of the contribution come not from differences in the reionization models – dn_b/dm is very similar in all three of these – but because of the different ionized gas densities during reionization.

Actually, the *duration* of reionization is the most important factor determining the amplitude of the patchy signal: the longer the contrast between ionized and neutral gas persists, the longer the patchiness continues to generate CMB fluctuations. The thick dotted curve illustrates this: it takes a model in which reionization lasts roughly twice as long as in the solid curve (but ends at the same time). Note how the patchy signal increases by nearly a factor of two. Thus, CMB secondary

anisotropies provide a tool to study not only the timing of reionization but also its duration.

The upper right set of curves in the right panel of Figure 12.2 show the sum of the Ostriker-Vishniac, patchy reionization, and primordial anisotropies (including lensing); the inset zooms in on the most useful part of the spectrum for studying the cosmic dawn. In this regime, the primordial anisotropies die off very quickly thanks to Silk damping inside the recombination surface; nevertheless, they are still very large near the peak of the patchy reionization contribution. Instead it will be the tail of this distribution, together with the cumulative Ostriker-Vishniac effect from all structure past reionization, that can be separated. The error bars here show estimates for ongoing ground-based CMB surveys; encouragingly, despite the apparent modest differences between the models, such an experiment can still easily distinguish them.

However, the primary challenge to making these measurements will be contaminants: both lower-redshift radio point sources and other CMB secondary anisotropies pose substantial problems. By far the strongest secondary at these angular scales is the thermal Sunyaev-Zel'dovich effect, which describes the frequency shift undergone by CMB photons scattering inside hot gas (in particular nearby galaxy clusters). In principle, because this scattering process is frequency-dependent (and disappearing entirely at 218 GHz). Constraints on the cosmic dawn from these secondaries will rely on accurate removal of these "foregrounds" as well as accurate modeling of the Ostriker-Vishniac effect at lower redshifts.

In principle, sources from the cosmic dawn may also contribute to this thermal Sunyaev-Zel'dovich signal, which arises when hot electrons inverse-Compton scatter off the CMB photons, transferring energy to the photon field. The magnitude of this effect is parameterized by the Compton-y parameter, which is the typical energy transfer per scattering times the number of scatterings. It also measures the resulting spectral distortion in a blackbody spectrum: in the Rayleigh-Jeans limit, the temperature distortion is $\Delta T/T = -2y$. The contribution from a volume element with electron scattering optical depth $d\tau_{\rm es}$ is

$$dy = \frac{k_B(T_e - T_\gamma)}{m_e c^2}\, d\tau_{\rm es}. \tag{12.7}$$

Thus, the total y distortion will depend upon both the ionization and thermal histories of the gas; importantly, however, it includes both gas *inside* and *outside* of galaxies.

The sources that reionized the Universe may very well make a non-negligible contribution to the overall y-distortion, due to the supernovae that inevitably accompany the massive stars able to ionize the IGM. Even though these supernovae may remain confined to the ISM of their host galaxies, they will still likely lose a substantial fraction of their energy to the CMB, inducing a y-distortion. Let us begin by computing the energy injection per baryon from such explosions. We write $\omega_{\rm SN}$ for the supernova energy produced per solar mass in stars formed; this is $\sim 10^{49}\,{\rm erg}/M_\odot$ for typical initial mass functions (IMFs) of stars. We then write the fraction of baryonic mass in stars, $f_\star f_{\rm coll}$ according to our usual definitions, in terms of the number of ionizing photons produced per baryon, Q. Then the

available thermal energy per baryon is

$$\frac{\epsilon_{\rm SN}}{\bar{n}_b} \sim 20Q \left(\frac{400}{N_\gamma f_{\rm esc}}\,\frac{\omega_{\rm SN}}{10^{49}\ {\rm erg}/M_\odot}\right)\ {\rm eV}. \tag{12.8}$$

Interestingly, this energy produced in supernovae is quite close to the amount actually needed to ionize the IGM (~ 13.6 eV per baryon), though the amount injected into the CMB may be much smaller.

Now let us assume that this energy injection occurs at some redshift $z_{\rm SN}$. The CMB spectral distortion will then satisfy

$$y = -\frac{1}{2}\frac{\Delta T_\gamma}{T_\gamma} \sim -\frac{1}{8}\frac{\Delta U_\gamma}{U_\gamma} \sim 10^{-6} f_{\rm comp}\left(\frac{10}{1+z_{\rm SN}}\right)\left(\frac{\epsilon_{\rm SN}/\bar{n}_b}{20\ {\rm eV}}\right) \tag{12.9}$$

where $f_{\rm comp}$ is the fraction of the supernova energy that is actually injected into the CMB. This is a reasonably large number: the current observational limit, from the FIRAS instrument on the *Cosmic Background Explorer* (COBE) is $y \leq 1.5 \times 10^{-5}$, and models predict signals from lower redshifts are just a factor of a few larger than equation (12.9). (On the other hand, the photoheating accompanying reionization does *not* make a substantial contribution: typically just a fraction of an eV is injected to the CMB per baryon, causing a very small signal compared to the much hotter gas at lower redshifts.)

However, such an observable signal hinges on the supernova energy being injected into the CMB, so that $f_{\rm comp} \sim 1$. Unfortunately, many other processes help to cool the remnants, especially if they remain confined to their galaxies. The most important is radiative cooling in the dense shell plowing through the IGM (or ISM). Simple estimates suggest that the energy contained in the explosion blastwave must be very large and that the remnant must spend the bulk of its time plowing through gas near the IGM density in order for more than just a few percent of the energy to be lost to the CMB. Three such alternatives are plausible: *(1)* powerful supernovae ($\sim 10^{53}$ erg) associated with the very massive, metal-free Population III stars that form in the earliest phases of structure formation [in the pair-instability mass range $130 < (M/M_\odot) < 260$] could release their energy into the IGM if their progenitor massive star photoevaporates the gas from their host halos; *(2)* the collective effects of many supernova could produce a starburst wind that escapes into the IGM; or *(3)* central quasar activity may generate a powerful outflow that removes the surrounding ISM before associated supernovae release most of their energy. The second scenario is certainly relaized in many starburst galaxies and is most likely responsible for the wide dispersal of metals through the IGM, but the detailed energetics of these winds and their importance for the CMB have not yet been studied in the regime in which Compton cooling may be significant.

If either of these mechanisms do cause a substantial y-distortion, the strong clustering of early stellar sources would also induce substantial angular fluctuations in the CMB temperature field – the *thermal Sunyaev-Zel'dovich* effect – which would then also be observable once the contribution from nearby hot galaxy clusters is subtracted. Again, the modeling of lower-redshift contaminants, which depends on uncertain factors like cluster cooling and turbulence, AGN feedback, and the poorly-understood properties of gas inside small galaxy groups, will be crucial to disentangling any possible high-redshift contribution.

12.2 DIFFUSE BACKGROUNDS FROM THE COSMIC DAWN

Although the many large telescopes planned for the next decades can study individual high-redshift galaxies and quasars in exquisite detail, one can also learn a great deal about these objects – at least statistically – by studying the *integrated* radiation backgrounds generated by such sources. If one does not attempt to identify individual galaxies, the telescope requirements are more modest, and the characteristics of the entire galaxy population can be measured.

Of course, these simplifications come with a price – diffuse backgrounds are much more difficult to interpret in the presence of other astronomical (or terrestrial) backgrounds. Much like the spin-flip background, the observational challenges are typically to extract the cosmological information from a large net signal. This task is easiest for a background generated by an emission line, because it will then have angular and frequency structure that clearly reflects the source population. Broadband backgrounds can only be distinguished by resolving other possible contaminants (as in the X-ray background discussed in §8.9.1); we will therefore focus on line backgrounds here.

The first interesting aspect of such backgrounds is their amplitude, which provides a measure of the total emissivity in this radiation mechanism through the cosmic dawn. For an emission line, where each observed frequency corresponds to a different distance, such a background is even more useful because it allows us to measure redshift evolution. However, one is always faced with the difficulty of separating the cosmic contribution at the relevant wavelength from any other mechanism. Sometimes this can be accomplished with a census of luminous sources (as in §8.9.1), but other times much more complicated methods are necessary (as in the spin-flip background, §11.4).

Let us assume that a source population has a comoving specific emissivity $\epsilon(\nu, \mathbf{r}, z)$ (with units of energy per time per frequency per comoving volume). Neglecting intervening absorption, the observed specific intensity I_ν at a frequency $\nu_{\rm obs}$ along a line of sight $\hat{\mathbf{n}}$ is the integral of the emissivity,

$$\nu_{\rm obs} I_{\nu_{\rm obs}} = \frac{c}{4\pi} \int dz\, \nu(z) \frac{\epsilon[\nu(z), \hat{\mathbf{n}} r(z), z]}{H(z)(1+z)^2}, \tag{12.10}$$

where $\nu(z) = \nu_{\rm obs}(1+z)$ is the emissionfreuqncy at a redshift z and $r(z)$ is the comoving distance along the line of sight to the source. If ϵ_ν extends over a wide frequency range, $I_{\nu_{\rm obs}}$ will therefore sample a wide redshift range. But if ϵ_i describes a line with rest wavelength ν_i, then the observed intensity will sample only a specific redshift $z_{\rm obs} = \nu_i/\nu_{\rm obs} - 1$. As with the spin-flip background, one can use this relation between observed frequency and distance to map the structure of the emissivity in three dimensions.

Without resolving individual sources, background measurements will be sensitive to large scale fluctuations in the emissivity, which we can easily parameterize with the galaxy power spectrum, either taken from simulations or computed with the halo model (see §3.7.1). For example, let us assume that the luminosity in line i of a halo with mass m is $L_i(m)$ and that a duty-cycle fraction $f_{\rm duty}$ (which may also be a function of mass) of dark matter halos with $m > m_{\rm min}$ emit in this line.

Then the mean comoving emissivity in the line is

$$\langle \epsilon_i \rangle (z) = \int_{m_{\rm min}}^{\infty} dm\, L_i(m) f_{\rm duty}(m) n(m). \quad (12.11)$$

Note that this has units of energy per (comoving) volume per second and is not a specific emissivity; rather, it includes all the emission in the line.

The fluctuations will also trace the population of massive halos, so it is natural to use the halo model to estimate them. However – for a diffuse background originating from unresolved sources – we are typically not concerned with the signal structure on scales below the typical halo's virial radius, which is anyway likely to be simply a set of point sources generated by each galaxy residing in the halo. We can therefore treat each halo as a single point source, ignoring the one-halo term.

In this case, the power spectrum of fractional emissivity fluctuations is $P_i(k) \approx P_i^{2h}(k)$, with

$$P_i^{2h}(k) = P_{\rm lin}(k) \left[\int dm\, \frac{L_i(m) f_{\rm duty}(m) n(m)}{\langle \epsilon_i \rangle} b_{\rm eff}(k|m), \right]^2 \quad (12.12)$$

where we have used the effective scale-dependent bias that incorporates non-linear biasing in the two-halo term (see §**??**) and assumed that we are on sufficiently large scales that $u_{\rm gal}(k|m) \approx 1$ for all the halos of interest. Mapping the spatial fluctuations in this background will therefore inform us about the number densities of these sources together with their scale-dependent bias; in principle, the latter is separable because of its special shape, so these different physical effects can be measured. (If the observations extend to fine enough scales to allow the one-halo term to be measured as well, they would allow a similar separation.)

Even though the one-halo term itself is only significant on small scales, the "shot noise" arising from the finite number of sources can be significant to these backgrounds, because halos massive enough to host galaxies are still relatively rare at these early times. This is especially true if only a subset of dark matter halos contribute to the background, either because of a small duty cycle or because only massive halos emit strongly in the relevant line. This shot noise term produces a white-noise spectrum (i.e., independent of k) with the amplitude of the emissivity power spectrum equal to

$$P_{{\rm shot},i}(k) = \int_{m_{\rm min}}^{\infty} dm\, \frac{L_i^2(m)}{\langle \epsilon_i \rangle^2} f_{\rm duty} n(m). \quad (12.13)$$

Note that $\langle \epsilon_i \rangle \propto \int dm L_i(m) f_{\rm duty}(m) n(m)$, so the shot noise term decreases with increasing source density, just as in the simple counting case (see §3.7.5).

Note that we have expressed these fluctuation spectra in their *fractional* forms; i.e. P_i has units of volume, independent of the emissivity. This means that they apply equally well to the emissivity or to the observed intensity – for example, one simply multiplies by the mean observed intensity to recover the latter in dimensional form.

12.2.1 The Near-Infrared Background

As an important example, we consider the integrated background from ultraviolet photons emitted during the cosmic dawn. This has two principal components: first,

a broadband background from stellar continua, and second, line and continuum components from the reprocessing of ionizing photons – principally Lyman-α . The latter is the most useful, because the broadband component involves a mix of many emission redshifts for a particular observed frequency.

To estimate the monopole amplitude of this background and its spectrum, we must estimate the contribution of all these different processes. Assuming that the background is generated entirely by star formation, a convenient parameterization is

$$\epsilon(\nu, z) = \dot{\rho}_\star(z) c^2 \sum_i \left\langle f_\nu^i \right\rangle , \tag{12.14}$$

where $\dot{\rho}_\star$ is the formation rate of stellar mass per comoving volume, the sum extends over all radiation processes (labeled by i), and $\left\langle f_\nu^i \right\rangle$ is the fraction of the stellar rest mass energy that is released in process i in the frequency interval $(\nu, \nu + d\nu)$ (for stars, it must therefore be averaged over both their initial mass function and their stellar lifetimes). The nuclear burning efficiency for stars implies, $\nu \left\langle f_\nu^i \right\rangle \sim 10^{-3}$. With this estimate, and assuming high redshifts for the integrand, equation (12.10) becomes

$$\nu_{\rm obs} I_{\nu_{\rm obs}} \approx 11 \ {\rm nW m^{-2} \ sr^{-1}} \int \frac{dz}{1+z} \dot{\rho}_\star(z) \left(\frac{10}{1+z} \right)^{5/2} \sum_i \frac{\nu(z) \left\langle f_{\nu(z)}^i \right\rangle}{10^{-3}}, \tag{12.15}$$

where $\dot{\rho_\star}$ is measured in units ${\rm M_\odot \ yr^{-1} \ Mpc^{-3}}$.

The processes contributing to this background are:

- The stellar continuum, which (below the Lyman edge) can be approximated as a blackbody with some effective temperature $T_{\rm eff}$ a function of stellar mass.
- Free-free and free-bound emission from H II regions. The total luminosity of this component may be written

  $$L_\nu^{\rm ff,fb} = \frac{\epsilon_\nu^{\rm ff,fb} \dot{Q}_i}{n_e n_p \alpha_B}, \tag{12.16}$$

  where $\epsilon_\nu^{\rm ff,fb}$ is the volume emissivities of these processes, $\dot{Q}_i$ is the production rate of hydrogen ionizing photons, and $\dot{Q}_i / n_e n_p \alpha_B$ is simply the volume of the ionized regions assuming the Stromgren sphere limit. Because these processes are generated by collisions of electrons and protons, the emissivity is

  $$\epsilon_\nu^{\rm ff,fb} = 4\pi n_e n_p g_{\rm eff} \frac{e^{-h\nu/k_B T}}{T}, \tag{12.17}$$

  where the "Gaunt factor" $g_{\rm eff} \sim 1$ depends weakly on temperature and density [43]. Thus, the total luminosity is independent of the local density. This means that, for the diffuse background we do not need to worry about whether the ionizing photons escape their galaxy or not – though that will clearly affect the small-scale spatial distribution of the photons, it has no effect on the overall luminosity, as long as the ionized regions are in ionization equilibrium.

- Recombination line emission – and in particular the Lyman-α line (higher Lyman-series photons are absorbed and cascade to either Lyman-α or the two-photon continuum discussed next). In §10, we saw that the Lyman-α profiles (both spatial and spectral) from individual galaxies depended strongly on the structure and physics of the galaxies and their local environments. However, the integrated emission is much simpler to estimate; assuming only that the mean free path of an ionizing photon is much shorter than the Hubble length, and that ionization equilibrium applies, the net emissivity of Lyman-α photons is simply (compare to equation 10.2)

$$\epsilon_{\rm Ly\alpha} = \frac{2}{3} T_{\rm dust} \frac{h\nu_\alpha}{\langle E_{\rm ion} \rangle} \epsilon_{\rm ion}, \tag{12.18}$$

  where $\epsilon_{\rm ion}$ is the emissivity of ionizing photons, $\epsilon_{\rm ion}/\langle E_{\rm ion} \rangle$ is the global ionization rate, and the factor of 2/3 accounts for those recombinations that do not produce a Lyman-α photon. Here $T_{\rm dust}$ accounts for absorption – and subsequent destruction – of Lyman-α photons by dust; note that there is no corresponding factor for absorption of the Lyman-α photons inside the IGM, because those photons are scattered but not destroyed. As with the free-free and free-bound processes, there is no distinction between ionizations that occur inside a galaxy and in the IGM, for both can produce Lyman-α photons through recombinations. We ignore the Balmer series or longer wavelength transitions because they carry much less energy than the Lyman-α line.

- Two-photon emission from (forbidden) decays between the $2s$ and $1s$ levels of hydrogen; the former level can be populated by radiative cascades following recombinations. The luminosity of this process is again proportional to the rate of ionizing photon production,

$$L_\nu^{2\gamma} = \frac{2h\nu}{\nu_\alpha} (1 - f_{\rm Ly\alpha}) P(\nu/\nu_\alpha) \dot{Q}_i, \tag{12.19}$$

  where $f_{\rm Ly\alpha} \approx 0.74$ is the fraction of cascades that result in Lyman-α photons (here we have assumed efficient mixing of the angular momentum states), the factor of two appears because of the two photons produced in each decay, and $P(x)dx$ is the normalized probability per decay of obtaining one photon in the range $dx = d\nu/\nu_\alpha$.

Figure 12.3 shows some example (monopole) spectra containing all of these processes; they are normalized to the overall star formation rate density $\dot{\rho}_\star$. These examples only include stars from $z = 7$–15, with $\dot{\rho}_\star$ held constant over that interval. The left and right panels vary the metallicity (which has a significant effect on the rate of ionizing photon production); within each panel, the different line styles take different prescriptions for the IMF. The solid line corresponds to a standard Salpeter IMF, with the number of stars forming per unit mass $\propto m^{-2.35}$. The dashed line refers to a Larson IMF, with the mass spectrum $\propto m^{-1}(1 + m/m_c)^{-1/35}$ where the characteristic mass is $m_c = 50\ M_\odot$. Finally, in the left-hand panel the double-dot-dashed curve takes a flat distribution by mass over the range 100–500 $M_\odot$.

The curves in each panel show the contributions of different processes to the overall spectrum. The uppermost is of course the total background per unit star

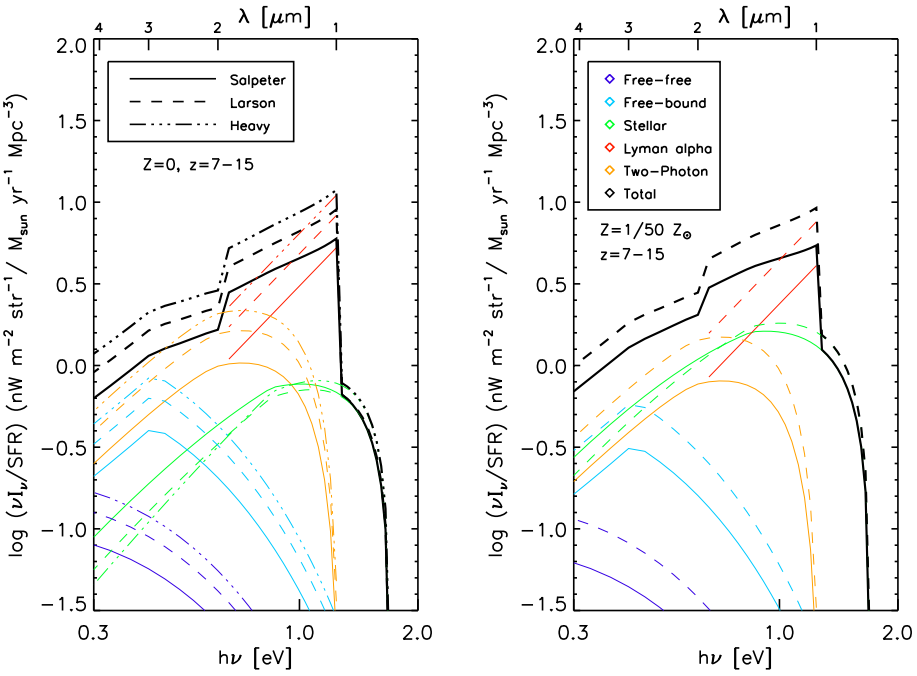

Figure 12.3 Near-infrared background spectrum. This example includes star formation from $z = 7$–15 (with constant comoving star formation rate), which dominates the background between 1–2 μm. The left panel assumes metal-free stars, while the right panel takes a metallicity $Z = 0.02 Z_\odot$. The different line styles show different IMFs (see text). The different curves show the total emission, the Lyman-α contribution, the stellar continua, the two-photon continua, the free-bound emission, and finally the free-free emission (from top to bottom at 1 μm). Figure credit: Fernandez, E. R. & Komatsu, E. 2006, ApJ, 646, 703.

formation rate. The straight line peaking at 1 μm shows the contribution from Lyman-α photons; note that the shape is simply a consequence of the assumption of a constant star formation rate and is not a robust prediction. The next process, which peaks near 1 μm as well, is the stellar continua; the curves here continue through the Lyman series without taking into account the sawtooth IGM absorption of these photons (see Fig. **??**). The curve peaking at somewhat lower energy, but with a comparable amplitude, comes from the two-photon decays: note that this is well below the Lyman-α peak not because significantly less energy goes into this process but because it is distributed over a wide frequency interval. Finally, the lowest amplitude curves show the contribution from free-bound and free-free emission, neither of which is significant.

Figure 12.3 shows several interesting points. First, in the most interesting wavelength range ($\sim$ 1–2 μm here) the background is usually dominated by the Lyman-α photons, especially for the (hotter) metal-free stars. This is because this line con-

tains a significant fraction of the entire ionizing luminosity of the starlight. Nevertheless, other processes provide non-trivial corrections at higher wavelengths. Note that stars at higher redshifts do not significantly affect the background in this band either, because their Lyman-α emission lies at longer wavelengths and (unless the comoving star formation rate becomes much higher at high redshifts) the stellar continua are quite weak. The near-infrared background therefore offers a relatively clean probe of the ionizing photon budget during the bulk of the cosmic dawn era.

Unfortunately, measuring this signal is extremely difficult: like the spin-flip background, this Lyman-α radiation suffers severe contamination from local sources – in addition to infrared emission from lower-redshift galaxies, the zodiacal light from our own solar system, which is roughly three times the expected signal. This background arises from dust particles that inhabit the ecliptic plane and scatter sunlight, and has proven very difficult to model with sufficient precision to extract the high-redshift signal reliably.

Fortunately, as with the spin-flip background, *fluctuations* in the near-infrared background light may be easier to detect than this monopole spectrum, because variations in the foreground contaminants have different spatial and spectral structure than the high-redshift light. In fact, the current best estimates of the monopole background come from measurements of the fluctuations themselves; they indicate that the excess over the known backgrounds cannot be much larger than $\sim 1\ \mathrm{nW m^{-2}\ sr^{-1}}$. This provides an interesting limit on the cosmic star formation rate at $z \sim 10$.

However, for higher-frequency diffuse backgrounds like this one, it is very difficult to recover fluctuations along the line of sight, as that requires an integral field spectrograph with sufficiently high spectral resolution to separate the features. To date, no such instruments are available over the wide fields of view necessary to measure this background in the near-infrared. Instead, these backgrounds are integrated over a finite frequency interval, and we typically measure the angular fluctuations (just as in the CMB) rather than the three-dimensional structure. This makes the foreground subtraction somewhat more difficult (unless several contiguous filters are used), because the spectral information is lost. Most often, the subtraction relies on modeling of the foreground structures.

In this case, we measure the band-averaged intensity $I(\hat{\mathbf{n}})$,

$$I(\hat{\mathbf{n}}) = \frac{c}{4\pi} \int dz\, \frac{\tilde{\epsilon}[\hat{\mathbf{n}} r(z), z]}{H(z)(1+z)^2}, \tag{12.20}$$

where $\tilde{\epsilon}$ is the integral of the comoving specific volume emissivity over the emission frequency range corresponding to the observed band.

To construct the angular fluctuation spectrum, we take the spherical harmonic transform of this quantity. Using Rayleigh's formula for the spherical decomposition of a plane wave,

$$e^{-i\mathbf{k}\cdot\mathbf{x}} = 4\pi \sum_{\ell m} (-i)^\ell j_\ell(kx) Y^*_{\ell m}(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{n}}), \tag{12.21}$$

where j_ℓ is the spherical Bessel function of order ℓ and $Y_{\ell m}$ are the spherical harmonics. We can then relate this transform to the three-dimensional Fourier trans-

form of the emissivity, $\tilde{\epsilon}(\mathbf{k}, z)$,

$$a_{\ell m} = c(-i)^\ell \int \frac{dz}{H(z)(1+z)^2} \int \frac{d^3\mathbf{k}}{(2\pi^3_)} \tilde{\epsilon}(\mathbf{k}, z) j_\ell[kr(z)] Y^*_{\ell m}(\hat{\mathbf{k}}). \qquad (12.22)$$

The angular power spectrum is usually expressed as the ensemble average of the spherical harmonic coefficients, $C_\ell = \left\langle |a_{\ell m}|^2 \right\rangle$. The ensemble average acts on the two factors of the (band-averaged) emissivity to give the three-dimensional power spectrum $P_{\tilde{\epsilon}}$, so

$$C_\ell = \frac{c^2}{8\pi^3} \int \frac{dz}{H(z)(1+z)^2} \int \frac{dz'}{H(z')(1+z')^2} \int k^2 dk P_{\tilde{\epsilon}}(k, z) j_\ell[kr(z)] j_\ell[kr(z')]. \qquad (12.23)$$

Fortunately, the inner integral can be simplified in the small-angle limit ($\ell \gg 1$). The integral of a product of two spherical Bessel functions is

$$\int k^2 dk\, j_\ell(kx) j_\ell(kx') = \frac{\pi}{2} \frac{\delta(x - x')}{x^2}. \qquad (12.24)$$

In the large ℓ limit, j_ℓ oscillates very rapidly, and the integral is dominated by $k \approx \ell/r$. Thus, provided that $P_{\tilde{\epsilon}}$ is a slowly varying function, we can pull it out of the integral and evaluate its argument at $k = \ell/r$. This **Limber's approximation** allows us to write

$$C_\ell \approx \left(\frac{c}{4\pi}\right)^2 \int \frac{dz}{H(z) r^2(z) (1+z)^4} P_{\tilde{\epsilon}} \left(k = \frac{\ell}{r(z)}, z\right). \qquad (12.25)$$

In other words, the angular power spectrum is simply the projection of the three-dimensional power spectrum, using the simple conversion $\ell \approx kr$. Note that, if the observed band is very thin, so that the line component of the background arises from only a narrow redshift window, one can easily invert the angular power spectrum to obtain the three-dimensional version, at least on scales larger than the width of the redshift window (on smaller scales the angular power is damped by cancellations along the line of sight). However, even with such a narrow window, the broadband component still arises from a wide range of redshifts, so it still requires modeling to invert properly.

The cumulative near-infrared background fluctuations will again be built from the Lyman-α emission, stellar continua, two-photon continua, and free-free/free-bound emission. However, unlike for the mean signal, the spatial location of this emission (within or outside of galaxies) is important, because recombinations that occur over the ~ 10 Mpc ionized bubbles in the IGM will change the spatial scales of the fluctuations (and create such low surface-brightness features that they will be all but unobservable in practice). In most circumstances, the fluctuations trace the galactic component and diminish as the escape fraction of UV photons, $f_{\rm esc}$ approaches unity.

Figure 12.4 shows some predictions for this fluctuating background in a range of models for star-forming galaxies during the reionization era. These examples all use a numerical simulation of reionization to construct the predicted signal; the simulation only fixes the *total* ionizing efficiency ζ, so there remains a good deal of freedom in the amplitude of the near-infrared background, which has a different

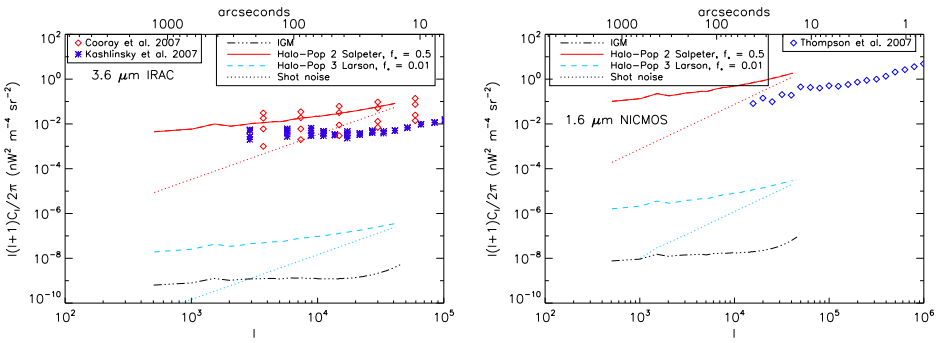

Figure 12.4 Comparison of recent observational data with theoretical predictions (taken from a numerical simulation of reionization) for the angular power spectrum of near-infrared background fluctuations. The two panels show results for the 3.6 μm IRAC band and the 1.6 μm NICMOS band (with observations taken from Cooray et al. 2007, Kashlinsky et al. 2007, and Thompson et al. 2007). In each one, the double-dot-dashed curve shows predictions for the IGM emission, while the solid and dashed curves show the range of predictions for the angular power spectra; these two limiting cases take a light, metal-poor population with high $f_\star$ and small $f_{\rm esc}$ (solid) and a heavy, metal-free population with low $f_\star$ and high $f_{\rm esc}$ (dashed curve), both normalized to have the same overall IGM ionization. The dotted curves show the associated shot noise terms. Figure credit: Fernandez et al. 2010, ApJ, 710, 1089.

dependence on $f_{\rm esc}$ in particular. The two models shown here span the range of possibilities in this simulation (though it is worth noting that there is even more variation outside of the particular star formation history in this simulation). The solid line corresponds to a low-metallicity stellar population with a Salpeter IMF (a mass spectrum $\propto m^{-2.35}$, ranging from 3–150 $M_\odot$), a very high $f_\star = 0.5$, and a relatively small $f_{\rm esc} = 0.19$. The dashed line, on the other hand, refers to very massive Population III stars (with the Larson IMF described previously and a characteristic mass $m_c = 250\,M_\odot$), a low $f_\star = 0.01$, and an escape fraction $f_{\rm esc} = 1$. Since the latter model has *no* line emission from the halo population, the fluctuations seen here are almost entirely due to the stellar continua.

The wide range of amplitudes for the angular power spectrum between these two extreme models illustrates how sensitive the near-infrared background is to the parameters of star formation in high-redshift galaxies; in combination with an independent measure of the IGM ionization state, this probe can help to break important degeneracies in the ionization efficiency. For a given ζ, the signal is maximized with a large star formation efficiency and small $f_{\rm esc}$ (so that more stars form, increasing both the stellar continuum and the overall production rate of ionizing photons *without* affecting the IGM) and a low-mass stellar population (which also boosts the stellar continuum as compared to the ionization rate).

The shapes of the power spectra are set by a combination of the two-halo term and the shot-noise contribution (which is shown separately by the dotted curves).

The latter is characterized by a white noise power spectrum with $C_\ell \propto \ell^0$. The former mirrors the linear power spectrum multiplied by the bias – in this case, nonlinear corrections are quite significant; otherwise the power spectrum would have actually *peaked* at $\ell \sim 10^3$. Instead, the two-halo term has $\ell(\ell+1)C_\ell \propto \ell^{1/2}$ when the nonlinear bias is included.

Interestingly, current observations are starting to pose interesting limits on these models: the two sets of points in the left hand panel show independent estimates of the residual fluctuations due to unresolved sources in images from the Infrared Array Camera on the *Spitzer* satellite at 3.6 μm; from our discussion of the mean background above, this is most sensitive to Lyman-α at very high redshifts and (most importantly in these models) continuum processes at a range of redshifts. The right panel shows data from the NICMOS camera on the *Hubble Space Telescope*, which operates at 1.6 μm and so is sensitive to Lyman-α emission at $z \sim 12$ (and continuum processes at all redshifts).

All of these efforts attempt to find fluctuations near the upper limits of theoretical expectations; however, note that the observed fluctuations may very well include any other unresolved population, such as faint low-redshift galaxies. In fact, the solid line in these models corresponds to a *mean* intensity of 15 and $60\ \mathrm{nWm^{-2}\ sr^{-1}}$ in these two bands, well above current limits, which suggests that in fact most of the observed fluctuations are due to lower-redshift or local contamination. The dashed line produces a mean background of 0.2–$0.8\ \mathrm{nWm^{-2}\ sr^{-1}}$ in these two bands, which can easily be accommodated by estimates of the mean intensity.

The dot-dashed curves in these plots show the prediction for the angular fluctuations generated inside the H II regions in the IGM. Given a particular simulation of reionization, this component is fixed by the densities and locations of the ionized bubbles. In any case it is very small, almost always negligible compared to the halo contribution, because of the very low IGM density (and hence recombination and collision rates).

In summary, the near-infrared background offers an intriguing view of the evolution of early stellar populations, highly complementary to other approaches which focus on measuring the ionization state of the IGM or on detecting individual bright objects. The primary challenges to understanding the background are, as in so many other areas, foregrounds: learning how to separate the relatively featureless angular and spectral behavior from other low-redshift contaminants and the zodiacal light. A combination with other, complementary probes may very well prove to be the best way to accomplish this goal.

12.2.2 Diffuse Backgrounds of Radio Lines

Another set of interesting diffuse backgrounds arise from radio and submillimeter lines: the same strong emission lines we have discussed in §**??**. These involve a molecule like CO, which has a forest of rotational lines with rest frequencies $J\nu_{\rm CO}$ for a transition from excited state J to $J-1$ (here $\nu_{\rm CO} = 115.3$ GHz), and the fine structure line of singly ionized carbon C II, which has a rest wavelength of 158 μm (or frequency 1.9 THz). Table 9.1 lists the most prominent interstellar emission

lines in star forming galaxies, along with their characteristic luminosity per star formation rate (in units of $L_\odot/(M_\odot/\mathrm{yr})$).

Some of these lines are particularly strong; for example, the C II line may carry as much as a percent of the total infrared luminosity of nearby quiescent galaxies like the Milky Way. Another advantage of these lower-frequency lines, which appear after redshifting in the centimeter or millimeter range, is that it is much easier to build instruments with both large fields of view and good spectral resolution to measure a diffuse background. Such measurements are then ideally suited for cross-correlation with other lines or the spin-flip background (see §12.3 below).

However, cross-correlation is likely necessary to recover any of these line backgrounds, because a single observed frequency will pick up emission from many different lines at many different redshifts. For example, an observed band around 30 GHz will be sensitive to CO(2-1) at $z = 6.7$ and to CO(1-0) at $z = 2.8$. One can isolate the high-redshift signal by comparing two *different* lines at the proper observed frequencies. For example, CO(1-0) at 10.5 GHz and CO(2-1) at 21 GHz both sample $z = 10$ galaxies. So long as no other lines have the same spacing, the cross-correlation between these two measurements will pick up information from $z = 10$ while eliminating all the contaminants, which simply contribute to the overall noise.

For unresolved point sources emitting in a pair of lines labeled 1 and 2, the cross power spectrum at a wavenumber k, can be approximated as

$$P_{1,2}(\mathbf{k}) = \bar{S}_1 \bar{S}_2 \bar{b}^2 P_{\rm lin}(\mathbf{k}) + P_{\rm shot}, \tag{12.26}$$

where $\bar{S}_1$ and $\bar{S}_2$ are the average fluxes in lines 1 and 2 respectively, $\bar{b}$ is the average bias factor of the sources, $P(\mathbf{k})$ is the matter power spectrum, and $P_{\rm shot}$ is the shot-noise power spectrum due to the discrete nature of galaxies (see §9.4). The root-mean-square error in the cross power spectrum at a particular k-mode is given by

$$\delta P_{1,2}^2 = \frac{1}{2}(P_{1,2}^2 + P_{1\rm total} P_{2\rm total}), \tag{12.27}$$

where $P_{1\rm total}$ and $P_{2\rm total}$ are the total power spectra corresponding to the first line and second line being cross correlated. Each of these includes a term for the power spectrum of contaminating lines, the target line, and detector noise. Figure 12.5 shows the expected errors in the determination of the cross power spectrum using the O I(63 μm) and O III(52 μm) lines at a redshift $z = 6$ for an optimized spectrometer on a 3.5 meter space-borne infrared telescope, providing background limited sensitivity for 100 diffraction-limited beams covering a square on the sky which is 1.7° across (corresponding to 250 comoving Mpc) and a redshift range of $\Delta z = 0.6$ (280 Mpc) with a spectral resolution of $(\Delta\nu/\nu) = 10^{-3}$ and a total integration time of 2×10^6 seconds.

Figure 12.6 shows another example of a diffuse background in the radio, the auto-correlation of CO(2-1). (Again, this would likely only be observable if cross-correlated with another CO line, but that only marginally affects the amplitude). The predictions are derived from a numerical simulation of reionization, and we show results for several different redshifts from the early to late stages of reionization (in the simulation, $Q_{\rm HII} = 0.82$ at $z = 6.8$ and 0.21 at $z = 8.8$).

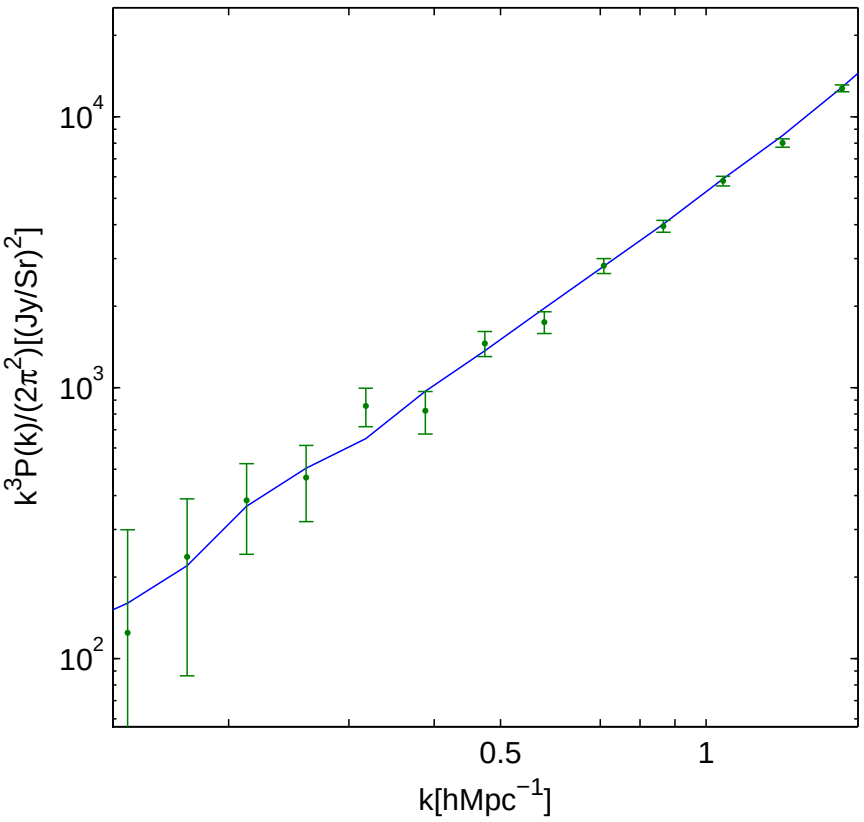

Figure 12.5 The cross power spectrum of OI(63 μm) and OIII(52 μm) at $z = 6$ measured from mock simulation data for a hypothetical infrared space telescope (see text). The solid line is the cross power spectrum measured when only line emission from galaxies in the target lines is included. The points are the recovered power spectrum when detector noise, contaminating line emission, galaxy continuum emission, and dust in our galaxy and the CMB are included. The error bars follow Eq. (12.27) with $P_{1\mathrm{total}}$ and $P_{2\mathrm{total}}$ calculated from the simulated data, including detector noise, contaminating line emission and sample variance. Figure credit: Visbal, E., Trac, H., & Loeb, A. *JCAP*, in press (2011).

The key assumptions to such a model are the mean intensity of the CO emission (which sets the overall normalization of the curves) and the luminosity-mass relationship of the source halos (which affects the shape of the curve by weighting different halos differently). The overall normalization requires two ingredients: an estimate of the total star formation rate density and a recipe for estimating the CO luminosity as a function of star formation rate (and possibly other factors, like the metallicity). Here, the latter is simply calibrated to local rapidly star-forming

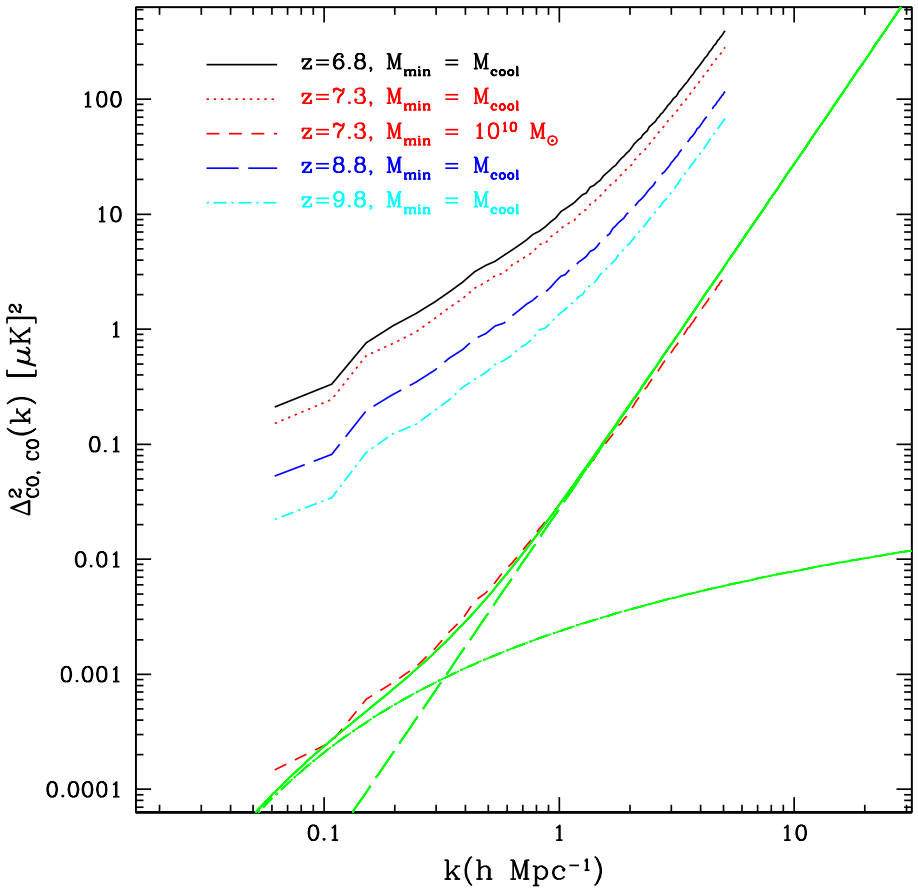

Figure 12.6 Three-dimensional power spectrum of CO(2-1) fluctuations in a simulation of reionization at several different epochs ($z = 6.8$, 7.3, 8.8, and 9.8). The CO emissivity of each galaxy is calibrated to local relations, and the total star formation rate density at $z = 6.8$ is fixed to the critical value required to maintain ionization (see text). The dashed line assumes that only massive galaxies emit CO radiation (possibly because the CMB dominates the dust temperature in smaller galaxies). For this scenario, we also show a simple model that includes shot noise (long-dashed curve) and an estimate of the two-halo term using the bias from the simulations (solid curve). Figure credit: Lidz A. et al., *ApJ*, in press (2011).

galaxies. As described in §**??**, this means that we assume that the local dust (and hence CO) excitation temperature is much larger than the CMB temperature even at these high redshifts and that the metallicity is near solar inside the molecular clouds.

The overall star formation rate density is fixed by requiring that, at $z = 6.8$, it is sufficient to *keep* the Universe ionized, according to equation (8.31). Note that, in the simulation, the Universe is not actually fully ionized at this time – but this nevertheless provides a useful and reasonable fiducial value. At higher redshifts, the simulation assumes that the star formation rate is proportional to the collapse fraction $f_{\rm coll}$.

As discussed in §**??**, the expected brightness of the CO lines depends on the detailed physics of the ISM of high-redshift galaxies, which is currently essentially unconstrained by observations. Thus, the overall amplitude of this signal is very uncertain. The short-dashed line in Figure 12.6 illustrates this with a model in which galaxies with $M < 10^{10}\,M_\odot$ are invisible in CO, possibly because their relatively small star formation rates are not enough to excite the gas temperature above the CMB temperature. In this case the signal declines by roughly a factor of 1000 on large scales, simply because of the drastically reduced overall emissivity of the gas.

Figure 12.6 reports the observed fluctuation amplitude in brightness temperature units, and shows that the typical value is $\sim 1\,\mu{\rm K}^2$ on moderately large scales. The shape of the power spectrum depends on both the large-scale clustering (or two-halo term in Eq. 12.12) and shot noise variations in the galaxy number counts (Eq. 12.13), with the former dominating on large scales and the latter on small scales. The Figure also shows the division between these two components for the case in which only massive galaxies emit; note that they provide a very good description of the signal. When less massive galaxies contribute to the CO emissivity, the shot noise term becomes less important (because the existence of many more sources implies smaller fractional fluctuations). However, the shape of the power spectrum remains nearly the same, because the nonlinear, scale-dependent bias is important for such galaxies.

While this signal is therefore thousands of times smaller than the spin-flip background, its appearnce at much higher frequencies (~ 10–50 GHz rather than ~ 50–200 MHz) means that the sky noise is also much, much smaller – in fact, at these frequencies the ~ 10 K noise inside the detectors dominates. Moreover, these frequencies are near those already used for CMB experiments, so this technology is well-developed, with both interferometers and large focal plane arrays available on the near-term horizon.

These properties enable CO mapping with much more modest instruments. Equation (11.56) suggests that a single, 3.5 m dish can reach a noise level of $\sim 1\,\mu$K per $10'$ pixel (its diffraction limit) in a spectral channel of a fractional width $\Delta\nu/\nu = 0.01$ at 30 GHz after an integration time of just a few days. Thus, a survey over several tens of square degrees could be accomplished relatively easily. In combination with detailed radio observations of individual galaxies from instruments like ALMA, such surveys would provide a complete census of molecular emission from the Cosmic Dawn.

12.3 THE CROSS-CORRELATION OF DIFFERENT PROBES

The diffuse backgrounds are relatively easy to measure in absolute terms, but often very difficult to isolate from the many other diffuse astrophysical signals that occupy the same wavebands. One way around this problem is to cross-correlate one measurement with another; such a correlation will eliminate any foreground contaminant that is not shared by both signals, greatly easing the extraction problem.

Moreover, such a cross-correlation can help to isolate interesting physical information. As an example, consider the cross-correlation of a galaxy survey with the spin-flip background. Ignoring redshift space distortions, the cross-power spectrum will be (c.f. eq. 11.53, recalling that the 21-cm signal is shaped by the baryon density and ionization fields)

$$\Delta^2_{21,g}(k) = T_0 \langle x_{\rm HI} \rangle [\Delta^2_{\delta,g}(k) + \Delta^2_{x,g}(k) + \Delta^2_{x\delta,g}(k)], \tag{12.28}$$

where the subscript g refers to the fractional overdensity in the galaxy field and x and δ denote the same for the neutral fraction and baryon density, respectively. On the relevant (large) scales, the first of these terms, $\Delta^2_{\delta,g}$, is simply proportional to the dark matter power spectrum multiplied by a bias factor (for the galaxy component). Similarly, the second term is simply the cross-correlation between dark matter density and ionization, multiplied by a bias factor. We have studied the behavior of this term in §8.8 and found two important results. First, on large scales, it is *negative* because the ionized bubbles appear where there are many galaxies, or conversely the gas is neutral only where there are no galaxies. Thus, the neutral gas and galaxy fields trace opposite ends of the dark matter density. On the other hand, it becomes small on scales below the bubble size, because the IGM bubble is entirely ionized regardless of its small-scale density structure (at least at the level probed by the spin-flip background). Importantly, there is a relatively sharp transition to zero in this field, though that is hidden in the spin-flip background itself by the other terms.

However, in the cross-correlation the turnover is more apparent, because the final term in equation (12.28) actually cancels $\Delta^2_{\delta,g}$ on small scales. To see this, note that the sum of these two terms is the Fourier transform of the quantity

$$\begin{aligned} x_{\rm HI}(1)\delta(1)n_g(2) &= \langle x_{\rm HI} \rangle [1 + \delta_x(1)]\delta(1)n_g(2) \\ &= \langle x_{\rm HI} \rangle \delta(1)n_g(2) + \langle x_{\rm HI} \rangle \delta_x(1)\delta(1)n_g(2), \end{aligned} \tag{12.29}$$

where the labels 1 and 2 refer to the two spatial positions and these two terms are respectively proportional to the Fourier transforms of $\Delta^2_{\delta,g}$ and $\Delta^2_{x\delta,g}$. However, we can explicitly write this two-point function:

$$\begin{aligned} \langle x_{\rm HI}(1)\delta(1)n_g(2) \rangle = \int & dx_{\rm HI}(1) d\delta(1) dn_g(2) x_{\rm HI}(1)\delta(1)n_g(2) \\ & \times P[x_{\rm HI}(1), \delta(1)|n_g(2)]P[n_g(2)], \end{aligned} \tag{12.30}$$

where we have simply expressed the correlation through a conditional probability function, and noted that the mean of y is the integral of y times its overall probability distribution. On separations much smaller than a typical bubble, the two points 1 and 2 will either be within the same ionized bubble or both neutral. In the former

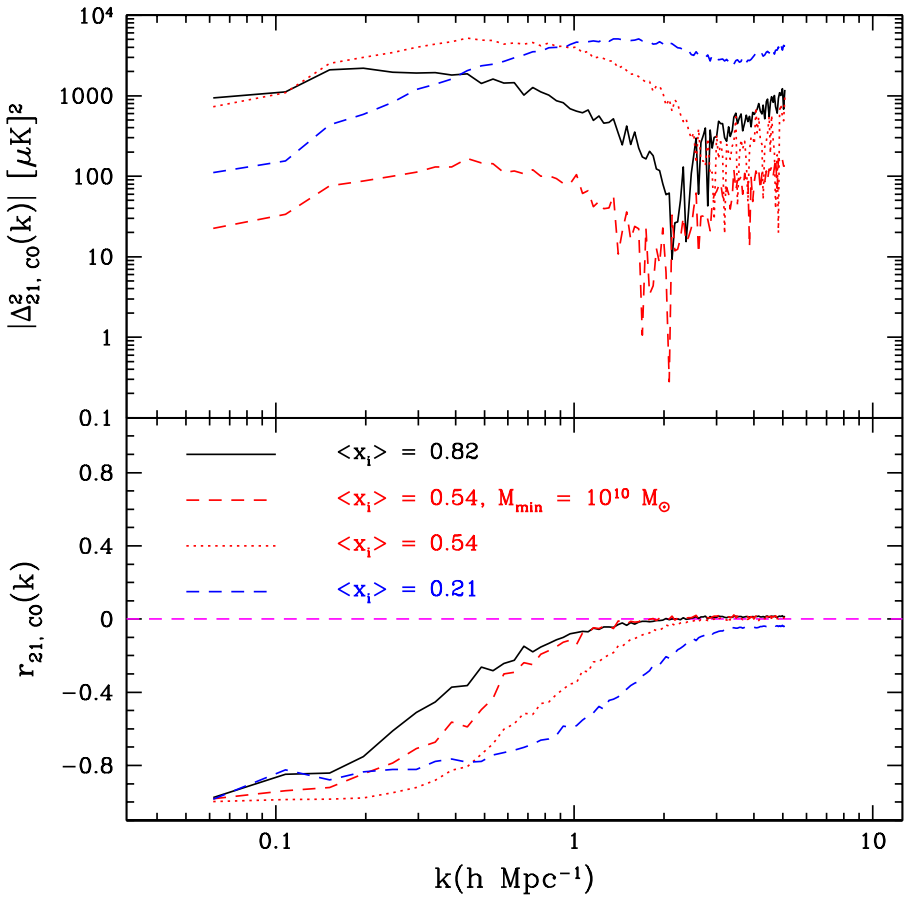

Figure 12.7 Cross-correlation between CO(2-1) emission and the spin-flip background in a numerical simulation of reionization (as in Fig. 12.6). The dot-dashed, dotted, and solid curves take $z = 9.8$, 7.3, and 6.8 (or $Q_{\rm HII} = 0.21$, 0.54, and 0.82 in this model), assuming that all galaxies emit CO(2-1). The long-dashed curve takes $z = 7.3$ but assumes that only massive galaxies emit CO. The top and bottom panels show the absolute value of the cross-power spectrum and the cross-correlation coefficient, respectively. Figure credit: Lidz, A. et al., *ApJ*, in press (2011).

case, $x_{\rm HI}(1) = 0$ and so the integral gets no contribution at all. In the latter case, outside of ionized bubbles, there must be no galaxies, so $n_g(2) = 0$. In this simple model, the cross-correlation therefore probes *only* the cross term $\Delta^2_{x,g}$, which contains a clear feature on the scale of the typical bubble.

One way to extract this information is therefore by conducting a galaxy redshift survey in the same volume as a 21-cm survey. However, such a project would be

very difficult, because the coarse resolution and huge fields of view of the radio telescopes (typically dozens or hundreds of square degrees, with several arcminute resolution) provide a very poor match to near-infrared galaxy surveys (which, at these high-redshifts, typically subtend at best several square arcminutes, but with exquisite angular resolution). A *diffuse* probe of the galaxy field therefore provides a much better match; in particular, the diffuse CO background is an excellent candidate because it also produces spectral fluctuations (note that broadband fluctuations, like the stellar continuum component of the near-infrared background, will not correlate as well because each wavelength comes from multiple redshifts).

Figure 12.7 shows the resulting cross-power spectrum for the CO(2-1) line in a numerical simulation of reionization in the top panel. The bottom panel shows the cross-correlation coefficient, which is defined as

$$r_{21,\mathrm{CO}}(k) = \frac{P_{21,\mathrm{CO}}(k)}{\sqrt{P_{21}(k)P_{\mathrm{CO}}(k)}}. \tag{12.31}$$

This quantity will be unity for perfectly correlated fields, equal to negative one for perfectly anti-correlated fields, and zero for uncorrelated fields.

The Figure shows predictions for three different stages of reionization. At every stage, the cross-power spectrum is negative on large-scales, reaching near perfect anti-correlation on sufficiently large scales. But this anti-correlation gradually turns into a nearly random association ($r = 0$) on small scales, with the turnover scale increasing as Q_{HII} increases, reflecting the rapidly growing ionized bubbles. This kind of cross-correlation therefore offers a clear measurement of the size of the H II bubbles, something that clearly affects the spin-flip power spectrum but that is much more difficult to extract from it in a model-independent manner.

Another interesting cross-correlation is between the CMB and the spin-flip background. On degree scales – much larger than the size of each ionized bubble – the cross-correlation is relatively easy to estimate. Qualitatively, a cross-correlation should arise because fluctuations in the density field source fluctuations in the ionized fraction (and hence 21-cm background) as well as in the baryon velocity field (which Doppler shifts CMB photons through scattering on free electrons). The Doppler contribution to the CMB usually cancels out, because photons traveling across an overdense region will be upscattered when they encounter gas falling toward the observer (on the far side of the overdensity) but be downscattered when they encounter gas falling away from the observer. However, if the ionized fraction changes across the overdensity, the cancellation will be imperfect.

If reionization were homogeneous, this would lead to an *anti-correlation* between the 21-cm brightness temperature and the CMB temperature. In this case, an overdensity would increase the 21-cm brightness but cool the CMB, as Thomson scattering would be more effective on the near side than the far side. But in the case of "inside-out" inhomogeneous reionization, we generically expect a *positive* correlation, because overdense regions host HII regions (decreasing the spin-flip signal) and still cool the CMB. Unfortunately, this correlation – which in principle provides a clean probe of the evolution of the average Q_{HII} – is still quite weak, with a cross-correlation coefficient $< 3\%$, because the primary CMB anisotropies dominate so strongly on the relevant scales (multipoles $\ell \sim 100$). Only in the case

that reionization occurs at very high redshifts ($z > 15$) will we be able to *detect* that there is a correlation, and even then at low confidence.

On smaller angular scales, the spin-flip background and CMB should anti-correlate in the case of inhomogeneous reionization, as the variations in the kSZ effect described in §12.1.2 appear where the 21-cm signal vanishes – on scales comparable to (or smaller than) the bubble size, the sign of the correlation is driven not by the velocity modes along the line of sight but by its projection $\mathbf{q}_\perp$ (see eq. 12.4). In principle, this small-scale correlation provides a clean probe of the ionized bubble properties; unfortunately, the cross-correlation is again dwarfed by the primary CMB anisotropies: although the kSZ signal itself peaks on quite large angular scales, the cross-correlation component is confined to $\ell < 8000$ (where the primary CMB signal is still large), largely because of cancellation of the structures in the integrated CMB map. Thus, although the CMB temperature-21 cm cross-correlation contains interesting physical information, it does not appear to be a useful observable in practice.

12.4 GRAVITATIONAL WAVES FROM BLACK HOLE MERGERS

As discussed in §7, when a galaxy forms – a small fraction of its gas typically settles to its center and makes (or feeds) a massive black hole there. Local galaxies show evidence for a central black hole mass that is roughly a fixed fraction ($\sim 0.1\%$) of the mass of their stellar spheroid. When two galaxies collide, their cores migrate by dynamical friction to the center of mass of the merged galaxy. If the galaxy is rich in gas (as expected for high-redshift galaxies), the orbit of the two black holes tightens on a timescale that is much shorter than the age of the Universe. The final phase of binary coalescence is driven by the emission of gravitational waves. The emitted waves could be detected by new observatories which are currently being planned or constructed.

As long as the binary separation is much larger than its Schwarzschild radius, the emitted gravitational wave luminosity can be derived in the post-Newtonian approximation. For two black holes on a circular orbit, the luminosity is

$$L_{\rm GW} = \frac{32}{5}\frac{G^4}{c^5}\frac{M^3\mu^2}{a^5}, \tag{12.32}$$

where a is the semi-major axis of the binary, $M = (M_1 + M_2)$ and $\mu = M_1 M_2/M$, with M_1 and M_2 being the masses of the binary members. As discussed in §7.4, the loss of energy to the emitted waves leads to a decrease in the binary separation a and an eventual coalescence of the two black holes over a time,

$$t_{\rm GW} = \frac{5}{256}\frac{c^5}{G^3}\frac{a^4}{M^2\mu}. \tag{12.33}$$

Supermassive binaries with comparable mass members merge in less than a Hubble time once their separation shrinks to $a < 10^{3.5} r_{\rm Sch}$ (where $r_{\rm Sch} = 2GM/c^2 = 3\times10^{11}{\rm cm}(M/10^6 M_\odot)$) or once their relative orbital velocity $v = (GM/a)^{1/2} > 10^{-2}c = 3\times10^3\ {\rm km\ s^{-1}}$.

Future detectors will be sensitive to the gravitational wave amplitude. To an order of magnitude, the observed wave amplitude from an equal mass binary with a Schwarzschild radius $r_{\rm Sch}$ and an orbital velocity v is given by, $h \sim (1+z)(r_{\rm Sch}/d_L)(v^2/c^2)$, where d_L is the luminosity distance to the binary. Since the signal amplitude only declines as (distance)$^{-1}$ rather than (distance)$^{-2}$ as for electromagnetic detectors which respond to photon flux, the first generation of sensitive gravitational wave observatories will already be able to find sources at cosmological distances.

More accurately, in a reference frame centered on the solar system's barycenter, the gravitational wave amplitude in its two polarization states is given by,

$$h_+ = \frac{2\mathcal{M}_z^{5/3}[\pi f_{\rm obs}]^{2/3}}{d_L}\left[1+(\hat{L}\cdot\hat{n})^2\right]\cos[2\Phi(t)]; \quad (12.34)$$

$$h_x = -\frac{4\mathcal{M}_z^{5/3}[\pi f]^{2/3}(\hat{L}\cdot\hat{n})}{d_L}\sin[2\Phi(t)]; \quad (12.35)$$

where the so-called "redshifted chirp mass" $\mathcal{M}_z \equiv (1+z)\mu^{3/5}/M^{2/5}$ sets the rate at which the binary shrinks, determining the "chirp" of their orbital frequency $P = 2\pi/\sqrt{GM/a^3}$. The precise orbital phase of the binary $\Phi(t)$ then depends on the masses and spins of the binary members, and yields the observed wave frequency, $f_{\rm obs}(t) = [\pi]^{-1}(d\Phi/dt)$, which is $(1+z)$ times smaller than the emitted wave frequency. The unit vector $\hat{n}$ points from the solar system frame to the binary – defining the sky coordinates of the source, and the unit vector $\hat{L}$ points along the binary angular momentum – defining the binary orientation relative to the line-of-sight. The inspiral signal does not provide explicitely the cosmological redshift separately from the binary masses, but the redshift can be inferred from $d_L(z)$ (or from an electromagnetic counterpart to the gravitational wave signal). Any particular detector is sensitive to a linear combination of the two polarization signals, with coefficients that depend on the orientation of the source relative to the detector.

The sensitivity of various gravitational wave observatories is shown in Figure 12.8. The Laser Interferometer Space Antenna (LISA[44]) is a planned space interferometer consisting of three spacecrafts whose positions mark the vertices of an equilateral triangle 5 km on a side in an orbit around the Sun. As evident from Figure 12.8, LISA will be able to detect $\sim 10^{4-7}M_\odot$ binaries out to arbitrary redshifts during the epoch of reionization. The next generation ground-based interferometer, Advanced-LIGO [45], will be sensitive to binaries involving black hole remnants of massive Pop-III stars (with $\sim 10^{2-3}M_\odot$) out to $z \gg 1$.

The expected event rate of massive binary mergers can be calculated based on the halo merger rate predicted by the excursion set formalism (§3.5.1) under various assumptions about the relation between the black hole and halo masses. For reasonable assumptions, LISA is expected to detect many cosmological events per year. The actual detection of these signals would open a new window into the Universe and enable to trace the hierarchical assembly of black holes in galaxies throughout cosmic history. Since gravitational waves pass freely through all forms of matter, gravitational wave observatories might discover new populations of black hole binaries that are electromagnetically faint because of their modest mass relative to bright quasars or because they are enshrouded in gas and dust.

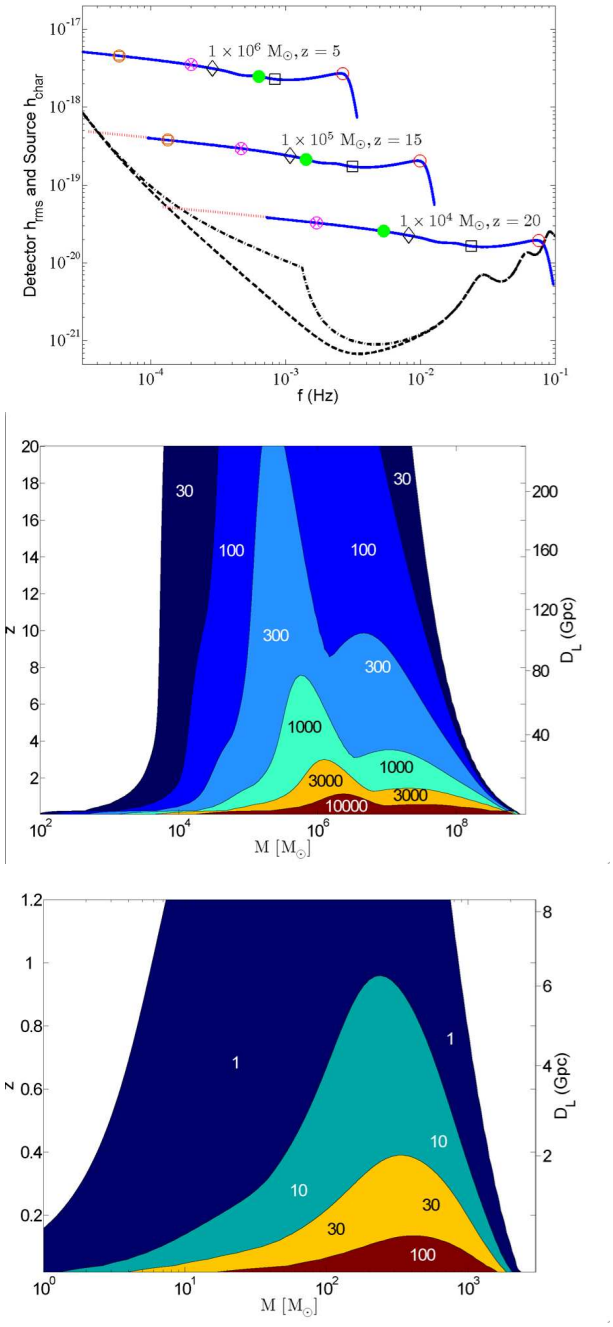

Figure 12.8 Sensitivity of the future gravitational wave observatories, LISA and Advanced-LIGO to equal-mass ($M_1 = M_2 = M/2$) binaries. *Top panel:* Root-mean-square noise amplitude of LISA h_{rms} from the detector only (dashed) and from the detector combined with the anticipated foreground confusion (dash-dotted), along with the characteristic amplitudes h_{char} of three binary masses M (solid). The locations on each h_{char} curve correspond to the peak amplitude (circle), 1 hour before the peak (filled circle), 1 day before the peak (circle with inscribed cross), and 1 month before the peak (circle with inscribed square) in the observer frame, as well as times of $25r_{\rm Sch}/c$ (square) and $500r_{\rm Sch}/c$ (diamond) before the peak in the source frame. *Middle panel:* Contour plot of the signal-to-noise ratio (SNR) with binary mass and redshift dependence for LISA. *Bottom panel:* SNR contour plot with mass and redshift dependence for Advanced-LIGO. Figure credit: Baker, J., et al. Phys. Rev. **D75**, 4024 (2007).

12.5 THE FOSSIL RECORD OF THE LOCAL GROUP

The focus of this book is on studies of the distant Universe: direct observations of galaxies (and the objects in them) or indirect probes of their environment. But of course these early generations of galaxies are the progenitors of every galaxy, including the Milky Way, and today's galaxies must contain remnants and signatures of these first structures. Here we will briefly discuss the prospects for *stellar archaeology* and its utility in understanding the Cosmic Dawn.

The hierarchical structure formation paradigm implies that the small dark matter halos in which the first stars and galaxies formed, merged to form larger and larger galaxies over time. During these violent merger events, existing gas reservoirs in galaxy cores would likely have undergone compression and formed stars, and continuing accretion at late times would have formed even more stars. Thus, the large majority of stars in today's galaxies had formed long after the Cosmic Dawn.

But what of the *existing* stars within the merging halos, that may have formed in pristine conditions (Population III) or shortly after, with very low metallicity? Recent large multi-object spectroscopic surveys have uncovered several hundred *extremely metal-poor* stars, characterized as having a relative abundance of iron to hydrogen at least three orders of magnitude smaller than that in the Sun (denoted as [Fe/H] < -3). Stellar archaeologists hope to use these stars to uncover information like the IMF and efficiency of second-generation star formation and the nucleosynthetic yields of the very first stars, which may have provided the heavy elements for the extremely metal-poor stars.

A related question is whether the small galaxies that surround the Milky Way – many containing so few stars that they remained hidden until the most recent generation of large surveys – can be traced back to the Cosmic Dawn. If so, they may carry imprints of important feedback, such as initial metal enrichment, the growth of the Lyman-Werner background, or reionization. For example, galaxies with halo masses $M \sim 10^8 M_\odot$ may have formed some stars before reionization and then been shut off by the photoheating that accompanied that event. Because these systems have undergone relatively little star formation in the past 10 Gyr, they may carry more obvious records of these events than large galaxies like the Milky Way.

12.5.1 Stellar Archaeology

Successfully using metal-poor stars to understand the earliest generations of structure require a number of inputs. First, one must *find* such stars – no easy task, only becoming possible with large spectroscopic surveys. Second, one must understand the dynamics of these stars and where they may have originated. Finally, one must relate their chemical abundances to the stars that enriched them, requiring detailed models of massive star supernovae, metal mixing in the ISM, and second-generation star formation.

- *Where are the extremely metal-poor stars?* So far, observations have found most of these stars in the outer halo of our Milky Way, that appears to have a

somewhat lower mean metallicity (centered at [Fe/H] ~ -2.2) than the inner halo (which is roughly three times more enriched).

Numerical simulations are consistent with this picture, showing that extremely metal-poor stars should appear throughout the Galaxy. This is largely because metal enrichment throughout the IGM is highly inhomogeneous, so that pockets of metal-free gas may persist until relatively late in the Universe's history, from $z \sim 5$–3. These relics would then be incorporated into the outer halo of the galaxy.

- *Where are the **oldest** extremely metal-poor stars?* If galaxies were composed solely of dark matter and stars, numerical simulations of hierarchical structure formation models would provide a fairly robust answer to this question: near the center of galaxies. These simulations show that galaxies form "inside-out," with the first objects to be accreted (i.e., the most overdense nearby regions, where the first stars would also have formed) residing closest to the bottom of the potential well of the galaxy, and later additions located farther and farther out in the halo of the galaxy. Thus, although extremely metal-poor stars may be spread throughout the halo, the *oldest* would be located near the center. This presents significant problems for surveys, as these few old stars would be buried in the much more numerous stars of our Galactic bulge and be subject to relatively large extinction.

  However, baryonic processes may mitigate this difficulty to some extent. In particular, spiral perturbations driven by accretion events can cause stars to migrate over large radial distances, getting deflected to much larger orbits. If so, such stars may be much more accessible to searches in the outskirts of galaxies, although their spatial distribution after a sequence of such events has not been well-quantified.

- *What are the "chemical fingerprints" of the first stars?* Once a set of such stars are found , stellar archaeologists hope to use their abundance patterns and other properties to learn about star formation in the early Universe – both at the time these stars formed and in the earlier generations that enriched them. The simplest approach is to use the "chemical fingerprints" present in these stars' abundance patterns to deduce the properties of the precursor stellar populations supernovae.

  Such efforts have a long tradition in astronomy, dating back to efforts to understand abundance patterns within our own solar system and in nearby stars. However, although the general problem is well-posed, extracting quantitative information remains difficult. Astrophysicists have a good qualitative understanding of the nuclear pathways through which heavy elements form – broadly, there are two different processes. In the *r-process*, which occurs during supernovae, neutrons are added to seed nuclei (usually ^{56}Ni) much more rapidly than β decay can occur. The resulting nuclei form a distinct pattern set by the locations of closed neutron shells, where the cross-section for continued neutron capture drops rapidly – however, the shells that form during the r-process are overabundant in neutrons and so suffer a sequence of

β decays before they reach stability. The contrasting *s-process* occurs when neutrons are added over long timescales, so that the nuclei can undergo β decay and grow through a sequence of stable nuclei. This occurs largely in the atmospheres of asymptotic giant-branch stars, over longer timescales than supernova nucleosynthesis.

Interestingly, the chemical patterns produced in supernovae depend strongly on the properties of their progenitors. As discussed in §5, stars of ~ 140–$260 M_{\odot}$ are subject to a pair-production instability, where much of the star's internal energy is lost when photon collisions create electron-positron pairs, locking up much of the thermal energy in the rest mass of those particles. The star then explodes violently. Before the explosion of pair-instability supernovae (PISN), the star has only a very small excess of neutrons, which strongly suppresses the formation of elements with an odd atomic number as compared to even ones. Moreover, these stars have large oxygen-fusing regions, which leads to an overproduction of Si and S compared to more normal supernovae.

Assuming that the extremely metal-poor stars were formed from gas enriched by only one or a few supernovae in the early generation of stars, one might therefore expect to see such fingerprints in their abundance patterns. However, to date this has proved not to be the case: instead, these stars –just like most of those in the Milky Way – appear to have been enriched by supernovae from stars with masses ~ 10–$40 M_{\odot}$, based largely on their overabundance of so-called α elements. These are nuclei made up of integer multiples of helium nuclei and are synthesized in the silicon-burning phase before such stars explode. There is thus so far little evidence for earlier generations of very massive ($> 100 M_{\odot}$) stars in these Galactic searches, although there are extragalactic clues for possible PISN explosions[46] and their odd-even abundance pattern in damped Lyman-α systems[47].

There are, however, several interesting anomalies that currently remain to be understood in the abundances of particular elements. One interesting feature is the large scatter in carbon and nitrogen, which can be greatly enhanced relative to iron. Many (but not all) of these stars also have enhanced s-process abundances; most likely, they have therefore been polluted by an intermediate-mass binary companion, which makes it difficult to tease out the abundances of the precursor supernovae.

The cooling rate by atomic carbon and oxygen dominates over that of molecular hydrogen once their abundances exceed $\sim 10^{-3.5}$ of the solar values. If such enhanced cooling is a pre-requisite for the ability of the gas to fragment into low-mass stars, then one would expect all low-mass stars in the Milky Way halo to show carbon or oxygen abundances above this threshold. Figure 12.9 shows that existing data is consistent with this theoretical expectation.

- *What can we learn about the IMF of the earliest stars?* The mere fact that no *metal-free* stars have been found – despite the relatively large number of extremely metal-poor stars – suggests that the very first generation of stars

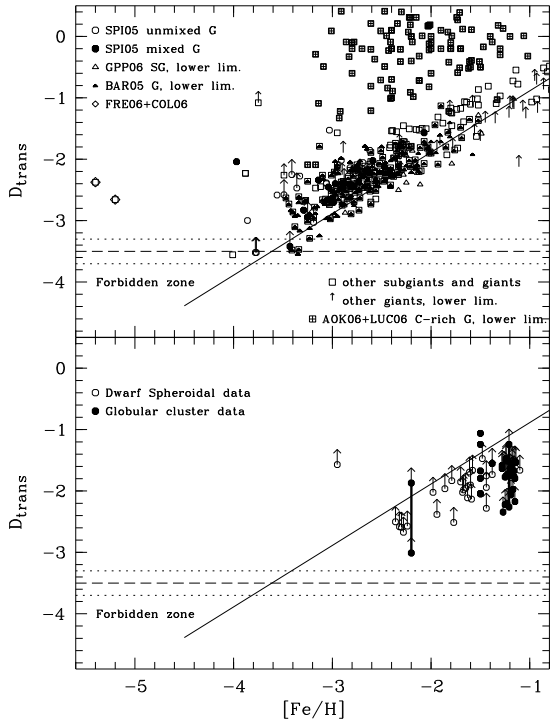

Figure 12.9 Measure of the carbon and oxygen abundance for metal-poor stars, $D_{\rm trans} \equiv \log_{10}\left(10^{\rm [C/H]} + 0.3 \times 10^{\rm [O/H]}\right)$ „ as a function of the iron abundance relative to hydrogen, [Fe/H]. *Top panel*: Galactic halo stars. *Bottom panel*: Stars in dwarf spheroidal galaxies and globular clusters. G and SG indicate giants and subgiants. The critical limit marked with a dashed line was predicted theoretically (Bromm, V., & Loeb, A. Nature **425**, 812 (2003)) by comparing the cooling rate for carbon and oxygen lines to H_2 cooling (which controls the formation of massive Pop III stars), predicting that no low mass stars should be found below this line. The dotted lines define the uncertainty in the theoretical prediction. Interestingly, all data points are above the theoretical line so far. Figure credit: Frebel, A., Johnson, J. L., & Bromm, V. *Mon. Not. R. Astron. Soc.* **380**, L49 (2007).

was skewed towards high-masses, with no evidence for Population III stars with a mass $< 0.8M_{\odot}$. The relatively common carbon-enhancement found in these stars also points to a higher characteristic mass, since it requires a large fraction of binary companions to have relatively large masses.

On the other hand, the lack of clear signatures of pair-instability supernovae, with masses $> 100M_{\odot}$, and the relative success of "normal" supernovae at reproducing the abundance patterns of extremely metal-poor stars, argues against very massive stars being common even in the earliest phases of structure formation – in contrast to the results of most numerical simulations. However, such an interpretation assumes that the heavy element products of the supernova mix efficiently with the ISM of the galaxy; if instead the mixing occurs only slowly, the second-generation stars may actually have relatively high metallicities and so lie outside the target area of existing surveys. A great deal remains to be learned from these surveys, and improved modeling of the transition from one stellar population to the next will be a crucial element of extracting the best possible information.

- *What additional information can we extract from these surveys?* It has become increasingly clear that, in the local Universe, star formation generally proceeds inside clusters of uniform abundance (albeit of widely varying masses). If similar processes occurred during the early generations of star formation, we would expect to find "clumps" in the abundance patterns characteristic of these star clusters, even if the stars themselves have dispersed on widely varying orbits. Such clumps could tell us about the environments in which early stars formed and the processes that regulated their birth; they would also provide "uniform" samples of stars over which one could average to study metal abundance across larger scales than individual stars, and with which one could account for anomalies like mass-transfer between binary companions.

12.5.2 Ultrafaint Dwarf Galaxies Around the Milky Way

Recent large spectroscopic surveys have revealed a wealth of information about the structure of our Milky Way galaxy and its immediate neighbors. In particular, they have enabled the discovery of a sizable class of "ultrafaint" dwarf galaxies, with a total luminosity $< 10^5 L_{\odot}$. These are particularly intriguing objects, because they appear to have undergone only one or a few star formation events (rather than the rich history common to larger galaxies), so the current stellar populations offer relatively clean tracers of the first generations of stars in these particular dark matter halos. Moreover, these galaxies are small enough that one can realistically imagine obtaining a fairly complete census of their stars.

The present data shows that these dwarf galaxies have low average metallicities, some as small as [Fe/H]~ -2.6, and substantial scatter in the abundances (at least an order of magnitude). Overall, the abundance patterns resemble those in the extremely metal-poor Galactic halo stars described in §12.5.1.

Thus, it seems that one of the feedback mechanisms that we have described ear-

lier was responsible for shutting down ongoing star formation in these galaxies. Three possibilities immediately come to mind: *(1)* ultraviolet feedback from the first stars that photodissociated H_2 and terminated cooling in minihalos below the atomic cooling threshold; *(2)* the supernova feedback generated by the first wave of star formation in the dwarf galaxy itself; and *(3)* photoheating from reionization (or possibly before) suppressing accretion onto these small halos, or possibly even evaporating any existing gas.

The canonical theoretical picture assumes that the initial burst of star formation in an isolated region contains only very high mass Population III stars. If the ultrafaint dwarfs – which contain low-mass stars – are then to form in the same dark matter halos, those galaxies must have been able to retain their gas (or at least reaccrete it) after these stars died. This generally argues against extremely low mass halos with shallow gravitational potential wells for the dwarfs. It also poses a challenge for any models that attribute the end of star formation in these galaxies to supernova winds: although such winds could have arisen from a later generation of star formation, the binding energy of the halo scales as $\propto M^2$, so one naively expects halos to become more stable as they grow.

The Lyman-Werner background on its own is also unlikely to be the factor stopping star formation in these galaxies. It is important only when H_2 dominates the cooling function; thus, because the stellar populations in these dwarfs have nonzero metallicity, it must not have been the dominant coolant in these populations (even if it was responsible for slowing down or stopping star formation in earlier phases).

The other key question is then how the first and second generations of star formation are related spatially. If minihalos can either retain or reacquire their gas after the initial burst of Pop III star formation, they may themselves have been able to form low-mass Pop II stars that survive to the present day. If, on the other hand, these halos instead lose their gas for a long period of time, larger objects – possibly above the atomic cooling threshold – would have held the bulk of the second generation of stars. Moreover, because these new stars form in a different halo than their Pop III progenitors, it is easier to imagine that they form in a burst mode that is able to evacuate the galaxy of its remaining gas, shutting down later star formation. Thus, understanding the dynamics and contents of these ultrafaint dwarfs offers the tantalizing possibility of constraining the large-scale pathways that enabled high-redshift star formation.

More detailed information is also available from these objects: for example, if an initial starburst did clear out the dwarf's gas, the stellar population should have little evidence of "self-enrichment:" rather, the stars might all have abundances characteristic of core-collapse (and possibly Pop III) supernovae, without any substantial s-process elements. The observed scatter in the metallicity within individual dwarfs also suggests that metals must not have been efficiently mixed across galactic scales, at least if the picture of a single burst of star formation is correct. Interestingly, old globular clusters have little or no apparent scatter in their abundances, implying much more efficient mixing in such systems.

The recent discovery of these dwarfs, and the rapidly increasing samples of metal-poor stars inside them and inside our own Galaxy, have opened a new win-

dow into studies of the impact of star formation in the early Universe. The implications of these studies are now only beginning to be understood, and a great deal of work on both the observational and theoretical ends is needed in order to disentangle the clues lying within. Stellar archaeology promises to remain a fascinating frontier for many years to come.

Appendix A

Notes

[1]de Bernardis, P., et al. *Nature* **404**, 955, (2000); Hanany, S., et al. *Astrophys. J.* **545**, L5 (2000); Miller, A. D., et al. *Astrophys. J.* **524**, L1 (1999).

[2]See, e.g. Peebles, P. J. E. *Principles of Physical Cosmology*, Princeton University Press (1993), in particular pages 62-65.

[3]See, e.g. overview in §8 of Barkana, R., & Loeb, A. *Phys. Rep.* **349**, 125 (2000); and also Haiman, Z., & Loeb, A. *Astrophys. J.* **483**, 21 (1997).

[4]For an overview of the current observational status, see Ellis, R. S. (2007), http://arxiv.org/abs/astro-ph/0701024.

[5]For advanced reading, see Mukhanov, V. *Physical Foundations of Cosmology*, Cambridge University Press, Cambridge (2005).

[6]http://public.web.cern.ch/public/en/LHC/LHC-en.html

[7]Loeb, A., Ferrara, A., & Ellis, R. S. *First Light in the Universe*, Saas-Fee Advanced Course **36**, Springer, New-York (2008), and references therein.

[8]Peebles, P. J. E. *Principles of Physical Cosmology*, Princeton University Press, (1993), p. 626.

[9]Eisenstein, D. J., & Hu, W. *Astrophys. J.* **511**, 5 (1999); Padmanabhan, T. *Theoretical Astrophysics, Volume III: Galaxies and Cosmology*, Cambridge University Press (2002), pp. 319-320.

[10]Haiman, Z., Thoul, A. A., & Loeb, A. *Astrophys. J.* **464**, 523 (1996).

[11]Barkana, R., & Loeb, A. *Astrophys. J.* **523**, 54 (1999).

[12]Loeb, A., & Zaldarriaga, M. *Phys. Rev.* **D71**, 103520 (2005).

[13]Gnedin, N. Y., & Hui, L. *Mon. Not. R. Astron. Soc.* **296**, 44 (1998). Note that we associate the filtering radius with π/k_F in analogy with the Jeans radius π/k_J; hence, our filtering mass M_F is $1/8$ of that defined by Gnedin & Hui who adopt a filtering radius of $2\pi/k_F$.

[14]Tseliakhovic, D. & Hirata, C., Phys. Rev. **D82**, 3520 (2010).

[15]Barkana, R., & Loeb, A. *Astrophys. J.* **531**, 613 (2000), and references therein.

[16]Bryan, G. & Norman, M., *Astrophys. J.* **495**, 80 (1998).

[17]Navarro, J. F., et al. *Mon. Not. R. Astron. Soc.* **402**, 21 (2010).

[18]Press, W. H., & Schechter, P. *Astrophys. J.* **187**, 425 (1974).

[19]Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N., *Astrophys. J.* **379**, 440 (1991); Lacey, C. G., & Cole, S. *Mon. Not. R. Astr. Soc.* **262**, 627 (1993).

[20]Reed, D. et al. *Mon. Not. R. Ast. Soc.* **374**, 2 (2007); Tinker, J. et al. *Astrophys. J.* **688**, 709 (2008).

[21]Peebles, P. J. E., *The Large-Scale Structure of the Universe*, Princeton University Press (1980).

[22]Zeldovich, Ya. B., *Astron. Astrophys.* **5**, 84 (1970).

[23]Shandarin, S.F., & Zel'dovich, Ya.B., *Rev. Mod. Phys.* **61**, 185 (1989).

[24]Scheuer, P. A. G. *Nature* **207**, 963 (1965).

[25]Gunn, J. E., & Peterson, B. A. *Astrophys. J.* **142**, 1633 (1965).

[26]Verner, D. A., Ferland, G. J., Korista, T., & Yakovlev, D. G. *Astrophys. J.* **465**, 487 (1996).

[27]Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **476**, 458 (1997).

[28]Strömgren, B. *Astrophys. J.* **89**, 526 (1939).

[29]Shapiro, P. R., & Giroux, M. L. *Astrophys. J.* **321** L107 (1987).

[30]Barkana, R., & Loeb, A. *Phys. Rep.* **349**, 129 (2001), and references therein.

[31]Wyithe, J. S. B., & Loeb, A. *Nature* **427**, 815 (2004); *Astrophys. J.* **610**, 117 (2004).

[32]mellema06

[33]Gnedin, N. *Astrophys. J.* **535**, 530 (2000).

[34]http://www.eso.org/sci/facilities/eelt/

[35]http://www.gmto.org/

[36]http://www.tmt.org/

[37]http://almaobservatory.org/
[38]Bradley, L., et al. *Astrophys. J.* **678**, 647 (2008).
[39]Stark, D., et al. *Astrophys. J.* **663**, 10 (2007); Bouwens, R. J., et al. *Astrophys. J* **690**, 1764 (2009).
[40]Wouthuysen, S. A. *Astron. J.* **57**, 31 (1952); Field, G. B. *Proc. IRE* **46**, 240 (1958).
[41]N. Kaiser, *Mon. Not. R. Astron. Soc.* **227**, 1 (1987).
[42]Ostriker, J. P., & Vishniac, E. T. *Astrophys. J.* **306**, L51 (1986); Vishniac, E. T. *Astrophys. J.* **322**, 597 (1987).
[43]Rybicki, G. B., & Lightman, A. P. Radiative Processes in Astrophysics, Wiley (1979), pp. 160-161.
[44]http://lisa.nasa.gov/
[45]http://www.advancedligo.mit.edu/
[46]Gal-Yam, A. et al. *Nature* **462**, 624 (2009).
[47]Cooke, R., et al. *Mon. Not. R. Astron. Soc.* **412**, 1047 (2011).

Appendix B

Recommended Further Reading

Cosmology

Padmanabhan, T., *Structure Formation in the Universe*, Cambridge University Press (1993)

Mukhanov, V., *Physical Foundations of Cosmology*, Cambridge University Press (2005)

Kolb, E. W., & Turner, M. S., *The Early Universe*, Addison Wesley (1990)

Peebles, P. J. E., *Principles of Physical Cosmology*, Princeton University Press (1993)

Loeb, A., *How Did the First Stars and Galaxies Form?*, Princeton University Press (2010)

Introduction to Astrophysics

Maoz, D., *Astrophysics in a Nutshell*, Princeton University Press (2007)

Schneider, P., *Extragalactic Astronomy and Cosmology*, Springer-Verlag (2006)

Radiative and Collisional Processes

Rybicki, G. B., & Lightman, A. P., *Radiative Processes in Astrophysics*, Wiley-Interscience (1979)

Osterbrock, D. E., & Ferland, G. J., *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (2nd edition), University Science Books (2006)

Compact Objects

Shapiro, S. L., & Teukolsky, S. A., *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects*, Wiley-Interscience (1983)

Peterson, B. M., *An Introduction to Active Galactic Nuclei*, Cambridge University Press (1997)

Galaxies

Binney, J., & Merrifield, M., *Galactic Astronomy*, Princeton University Press (1998)

Binney, J., & Tremaine, S., *Galactic Dynamics* (2nd edition), Princeton University Press (2008)

Appendix C

Useful Numbers

Fundamental Constants		
Newton's constant (G)	=	$6.67 \times 10^{-8}\ \mathrm{cm^3\ g^{-1}\ s^{-2}}$
Speed of light (c)	=	$3.00 \times 10^{10}\ \mathrm{cm\ s^{-1}}$
Planck's constant (h)	=	$6.63 \times 10^{-27}\ \mathrm{erg\ s}$
Electron mass (m_e)	=	$9.11 \times 10^{-28}\ \mathrm{g} \equiv 511\ \mathrm{keV}/c^2$
Electron charge (e)	=	$4.80 \times 10^{-10}\mathrm{esu}$
Proton mass (m_p)	=	$1.67 \times 10^{-24}\ \mathrm{g} = 938.3\ \mathrm{MeV}/c^2$
Boltzmann's constant (k_B)	=	$1.38 \times 10^{-16}\ \mathrm{erg\ K^{-1}}$
Stefan-Boltzmann constant (σ)	=	$5.67 \times 10^{-5}\ \mathrm{erg\ cm^{-2}\ s^{-1}\ K^{-4}}$
Radiation constant (a)	=	$7.56 \times 10^{-15}\ \mathrm{erg\ cm^{-3}\ K^{-4}}$
Thomson cross-section (σ_T)	=	$6.65 \times 10^{-25}\ \mathrm{cm^2}$
Astrophysical numbers		
Solar mass ($M_\odot$)	=	$1.99 \times 10^{33}\ \mathrm{g}$
Solar radius ($R_\odot$)	=	$6.96 \times 10^{10}\ \mathrm{cm}$
Solar luminosity ($L_\odot$)	=	$3.9 \times 10^{33}\ \mathrm{erg\ s^{-1}}$
Hubble constant today (H_0)	=	$100h\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$
Hubble time (H_0^{-1})	=	$3.09 \times 10^{17} h^{-1}\ \mathrm{s} = 9.77 \times 10^9 h^{-1}\ \mathrm{yr} \equiv 3h^{-1}\ \mathrm{Gpc}/c$
critical density (ρ_c)	=	$1.88 \times 10^{-29} h^2\ \mathrm{g\ cm^{-3}} = 1.13 \times 10^{-5} h^2 m_p \mathrm{cm^{-3}}$
Unit conversions		
1 parsec (pc)	=	$3.086 \times 10^{18}\ \mathrm{cm}$
1 kilo-parsec (kpc)	=	$10^3\ \mathrm{pc}$
1 mega-parsec (Mpc)	=	$10^6\ \mathrm{pc}$
1 giga-parsec (Gpc)	=	$10^9\ \mathrm{pc}$
1 Astronomical unit (AU)	=	$1.5 \times 10^{13}\ \mathrm{cm}$
1 year (yr)	=	$3.16 \times 10^7\ \mathrm{s}$
1 light year (ly)	=	$9.46 \times 10^{17}\ \mathrm{cm}$
1 eV	=	$1.60 \times 10^{-12}\ \mathrm{ergs} \equiv 11,604\ \mathrm{K} \times k_B$
1 erg	=	$10^{-7}\ \mathrm{J}$
Photon wavelength ($\lambda = c/\nu$)	=	$1.24 \times 10^{-4}\ \mathrm{cm\ (photon\ energy/1\ eV)^{-1}}$
1 nano-Jansky (nJy)	=	$10^{-32}\ \mathrm{erg\ cm^{-2}\ s^{-1}\ Hz^{-1}}$
1 Angstrom (Å)	=	$10^{-8}\mathrm{cm}$
1 micron (μm)	=	$10^{-4}\mathrm{cm}$
$1\ \mathrm{km\ s^{-1}}$	=	1.02 pc per million years
1 arcsecond ($''$)	=	4.85×10^{-6} radians
1 arcminute ($'$)	=	$60''$
1 degree ($^\circ$)	=	$3.6 \times 10^{3''}$
1 radian	=	57.3°

Appendix D

Glossary

- **Baryons**: strongly interacting particles made of three quarks, such as the proton and the neutron from which atomic nuclei are made. Baryons carry most of the mass of ordinary matter, since the proton and neutron masses are nearly two thousand times higher than the electron mass. Electrons and neutrinos are called **leptons** and are only subject to the electromagnetic, gravitational and weak interactions.

- **Big Bang**: the moment in time when the expansion of the Universe started. We cannot reliably extrapolate our history before the Big Bang because the densities of matter and radiation diverge at that time. A transition through the Big Bang could only be described by a future theory that will unify quantum mechanics and gravity.

- **Blackbody radiation**: the radiation obtained in complete thermal equilibrium with matter of some fixed temperature. The intensity of the radiation as a function of photon wavelength is prescribed by the Planck spectrum. The best experimental confirmation of this spectrum was obtained by the COBE satellite measurement of the Cosmic Microwave Background (CMB).

- **Black hole**: a region surrounded by an **event horizon** from where no particle (including light) can escape. A black hole is the end product from the complete gravitational collapse of a material object, such as a massive star or a gas cloud. It is characterized only by its mass, charge, and spin (similarly to elementary particles).

- **Cosmology**: the scientific study of the properties and history of the Universe. This research area includes **observational** and **theoretical** sub-fields.

- **Cosmic inflation**: an early phase transition during which the cosmic expansion accelerated, and the large-scale conditions of the present-day Universe were produced. These conditions include the large-scale homogeneity and isotropy, the flat global geometry, and the spectrum of the initial density fluctuations, which were all measured with exquisite precision over the past two decades.

- **Cosmic Microwave Background (CMB)**: the relic thermal radiation left over from the opaque hot state of the Universe before cosmological recombination.

- **Cosmological constant (dark energy)**: the mass (energy) density of the vacuum (after all forms of matter or radiation are removed). This constituent introduces a repulsive gravitational force that accelerates the cosmic expansion. The cosmic mass budget is observed to be dominated by this component at the present time (as it carries more than twice the combined mass density of ordinary matter and dark matter).

- **Cosmological principle**: a combination of two constraints which describe the Universe on large scales: *(i)* homogeneity (same conditions everywhere), and *(ii)* isotropy (same conditions in all directions).

- **Dark matter**: a mysterious dark component of matter which only reveals its existence through its gravitational influence and leaves no other clue about its nature. The nature of the dark matter is unknown, but searches are underway for an associated weakly-interacting particle.

- **Gamma-Ray Burst (GRB)**: a brief flash of high-energy photons which is often followed by an afterglow of lower energy photons on longer timescales. Long-duration GRBs (lasting more than a few seconds) are believed to originate from relativistic jets which are produced by a black hole after the gravitational collapse of the core of a massive star. They are often followed by a rare (Type Ib/c) supernova associated with the explosion of the parent star. Short duration GRBs are thought to originate also from the coalescence of compact binaries which include two neutron stars or a neutron star and a black hole.

- **Hubble parameter** $H(t)$: the ratio between the cosmic expansion speed and distance within a small region in a homogeneous and isotropic Universe. Formulated empirically by Edwin Hubble in 1929 based on local observations of galaxies. H is time dependent but spatially constant at any given time. The inverse of the Hubble parameter, also called the **Hubble time**, is of order the age of the Universe.

- **Hydrogen**: a proton and an electron bound together by their mutual electric force. Hydrogen is the most abundant element in the Universe (accounting for $\sim 76\%$ of the primordial mass budget of ordinary matter), followed by helium ($\sim 24\%$), and small amounts of other elements.

- **Jeans mass**: the minimum mass of a gas cloud required in order for its attractive gravitational force to overcome the repulsive pressure force of the gas. First formulated by the physicist James Jeans.

- **Galaxy**: an object consisting of a luminous core made of stars or cold gas surrounded by an extended halo of dark matter. The stars in galaxies are often organized in either a disk (often with spiral arms) or ellipsoidal configurations, giving rise to **disk** (spiral) or **elliptical** (spheroidal) galaxies, respectively. Our own Milky Way galaxy is a disk galaxy with a central spheroid. Since we observe our Galaxy from within, its disk stars appear to cover a strip across the sky.

- **Linear perturbation theory**: a theory describing the gravitational growth of small-amplitude perturbations in the cosmic matter density, by expanding the fundamental dynamical equations to leading order in the perturbation amplitude.

- **Lyman-α transition**: a transition between the ground state ($n = 1$) and the first excited level ($n = 2$) of the hydrogen atom. The associated photon wavelength is 1216Å.

- **Neutron star**: a star made almost exclusively of neutrons, formed as a result of the gravitational collapse of the core of a massive star progenitor. A neutron star has a mass comparable to that of the Sun and a mass density comparable to that of an atomic nucleus.

- **Quasar**: a bright compact source of radiation which is powered by the accretion of gas onto a massive black hole. The relic (dormant) black holes from quasar activity at early cosmic times are found at the centers of present-day galaxies.

- **Recombination of hydrogen**: the assembly of hydrogen atoms out of free electrons and protons. Cosmologically, this process occurred 0.4 million years after the Big Bang at a redshift of $\sim 1.1 \times 10^3$ when the temperature first dipped below $\sim 3 \times 10^3$ K.

- **Reionization of hydrogen**: the break-up of hydrogen atoms, left over from cosmological recombination, into their constituent electrons and protons. This process took place hundreds of millions of years after the Big Bang, and is believed to have resulted from the UV emission by stars in the earliest generation of galaxies.

- **Supernova**: the explosion of a massive star after its core consumed its nuclear fuel.

- **21-cm transition**: a transition between the two states (up or down) of the electron spin relative to the proton spin in a hydrogen atom. The associated photon wavelength is 21 cm.

- **Star**: a dense, hot ball of gas held together by gravity and powered by nuclear fusion reactions. The closest example is the Sun.