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.

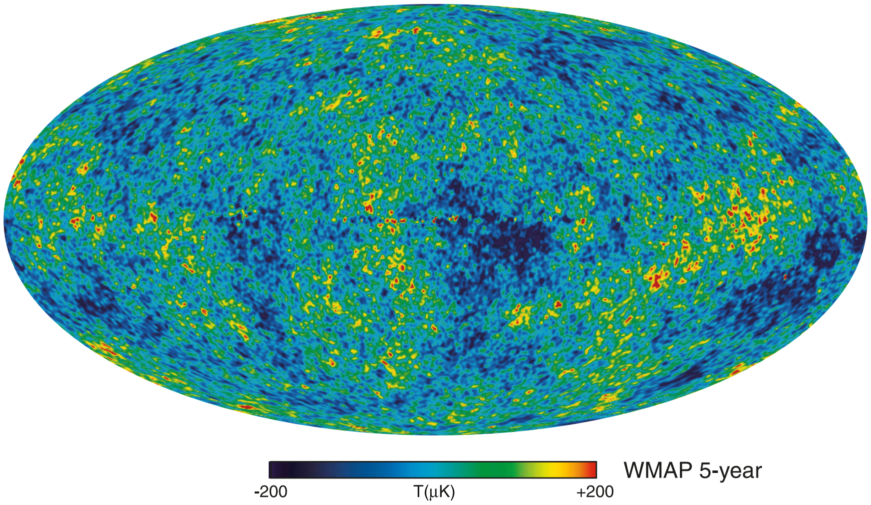

Figure 1.1 Image of the Universe when it first became transparent, 400 thousand years after the Big Bang, taken over five years by the *Wilkinson Microwave Anisotropy Probe* (WMAP) satellite (http://map.gsfc.nasa.gov/). Slight density inhomogeneities at the level of one part in $\sim 10^5$ in the otherwise 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

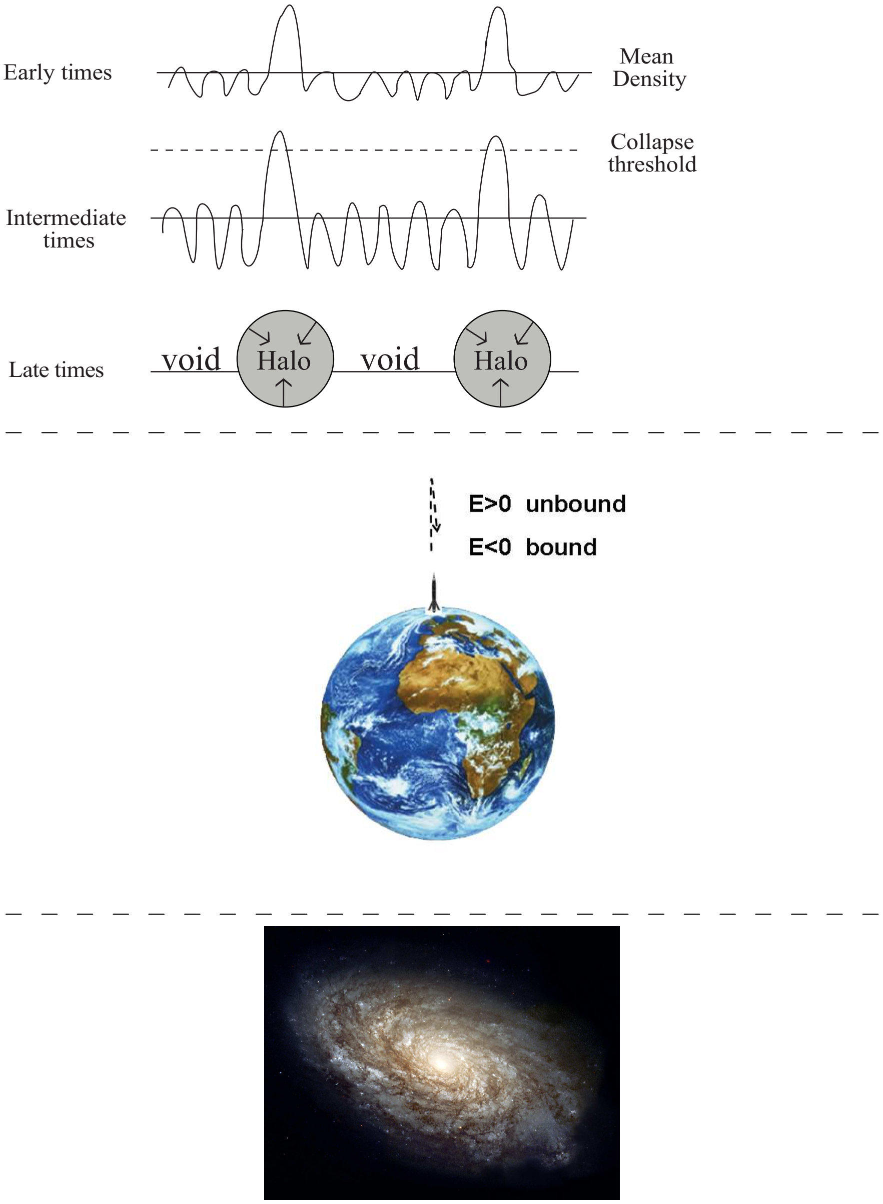

Figure 1.2 *Top:* Schematic illustration of the growth of perturbations to collapsed halos through gravitational instability. Once the overdense regions exceed a threshold density contrast above unity, they turn around and collapse to form halos. The material that makes the halos originated in the voids 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,

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

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,

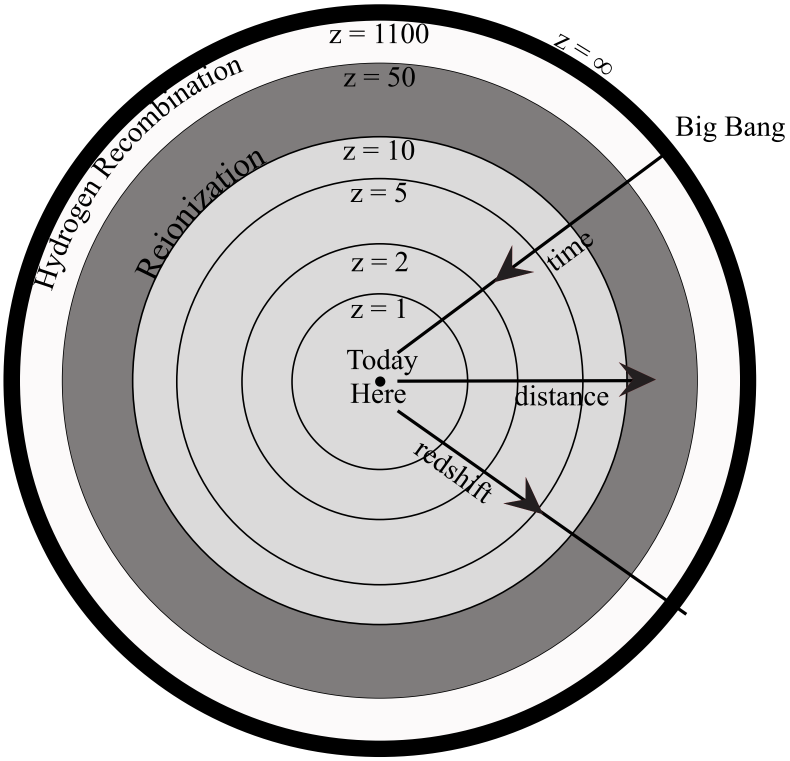

Figure 1.3 Cosmic archaeology of the observable volume of the Universe, in comoving coordinates (which factor out the cosmic expansion). The outermost observable boundary ($z = \infty$) marks the comoving distance that light has traveled since the Big Bang. Future observatories aim to map most of the 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).

1.3 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

$$\frac{d^2a}{dt^2} = -\frac{GM_{\rm grav}}{a^2}. \tag{1.4}$$

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 $-3pa\dot{a}/c^2 = a^2\dot{\rho} + 3\rho a\dot{a}$,

[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.

where a dot denotes a time derivative. Multiplying equation (1.4) by $\dot{a}$ and making use of this relation yields our familiar result

$$E = \frac{1}{2}\dot{a}^2 - \frac{GM}{a}, \tag{1.5}$$

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 (1.5) can be re-written as

$$\frac{E}{\frac{1}{2}\dot{a}^2} = 1 - \Omega, \tag{1.6}$$

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

$$\rho_c = \frac{3H^2}{8\pi G} = 9.2 \times 10^{-30}\frac{\mathrm{g}}{\mathrm{cm}^3}\left(\frac{H}{70\ \mathrm{km\ s^{-1}Mpc^{-1}}}\right)^2. \tag{1.7}$$

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

$$\frac{H(t)}{H_0} = \left[\frac{\Omega_m}{a^3} + \Omega_\Lambda + \frac{\Omega_r}{a^4}\right]^{1/2}, \tag{1.8}$$

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 $a \propto \exp\{H_{\mathrm{vac}}t\}$ with $H_{\mathrm{vac}} = (8\pi G\rho_{\mathrm{vac}}/3)^{1/2}$ 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 ρ_{grav} already at $z < 0.7$ or six billion years ago. Figure 1.5 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 ($1 < z < 10^3$),

$$t \approx \frac{2}{3H_0\Omega_m{}^{1/2}(1+z)^{3/2}} = \frac{0.95 \times 10^9\ \mathrm{years}}{[(1+z)/7]^{3/2}}. \tag{1.9}$$

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.[3] 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.[vii] 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

[vii] 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.

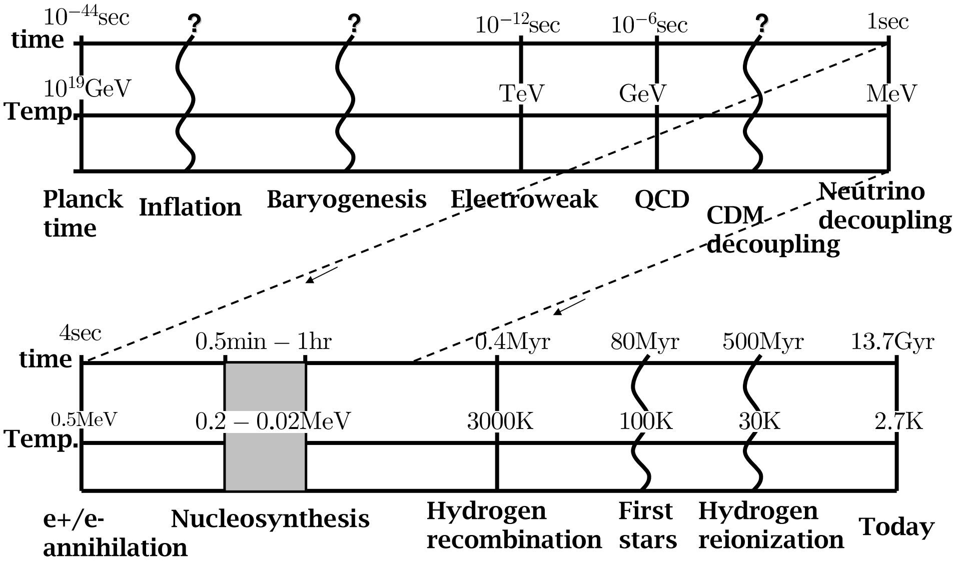

Figure 1.4 Following inflation, the Universe went through several other milestones which left a detectable record. These include baryogenesis (which resulted in the observed asymmetry between matter and anti-matter), the electroweak phase transition (during which the symmetry between electromagnetic 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.

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"[viii] $[(hc^5/G)^{1/2} \sim 10^{19}\ \mathrm{GeV}]$, 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.

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 1.5). 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

[viii] 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.

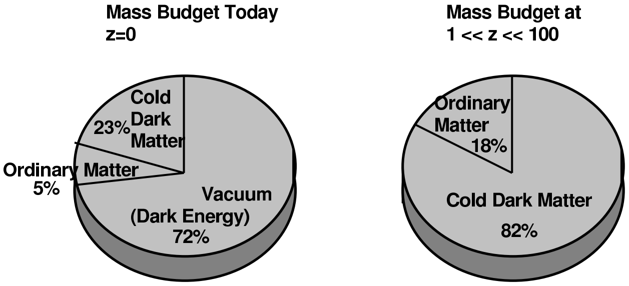

Figure 1.5 Mass budgets of different components in the present day Universe and in the infant Universe when the first galaxies formed (redshifts $z = 10$–50). The CMB radiation (not shown) makes up a fraction $\sim\ 0.03\%$ of the budget today, but was dominant at redshifts $z\ >\ 3,300$. The cosmological constant (vacuum) contribution was negligible at high redshifts ($z \gg 1$).

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 produced in small quantities through collisions of energetic particles in future laboratory experiments such as the Large Hadron Collider (LHC).[4] 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 g alaxies. 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

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

conditions. Below we follow the process through which late time complexity appeared and established an irreversible arrow to the flow of cosmic time.[ix]

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 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.

[ix] 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.

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.[5]

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 (1.6) 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 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 ($D(t)$ →constant) as their growth is suppressed by the accelerated expansion of space.

The initial perturbation amplitude varies with spatial scale. 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 captured by expressing 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 $\delta_{\mathbf{k}} = \int d^3x\, \delta(x) e^{-i\mathbf{k}\cdot\mathbf{x}}$, 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. 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 the fluctuations are determined by the variance of the different $\mathbf{k}$-modes given by the so-called power spectrum, $P(k) = (2\pi)^{-3} \left\langle |\delta_{\mathbf{k}}|^2 \right\rangle$, where the angular brackets denote an average over the entire statistical ensemble of modes.

In the standard cosmological model, inflation produces a 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 distance traveled by light during the age of the Universe), and so this spectrum is called "scale-invariant."[ii] The growth of perturbations in a CDM Universe results in a modified final power spectrum characterized by a turnover at a scale of order the horizon cH^{-1} at matter-radiation equality, and a small-scale asymptotic shape of $P(k) \propto k^{n_s-4}$. The turnover results from the fact that density perturbations experience almost no growth during the radiation dominated era, because the Jeans length then ($\sim ct/\sqrt{3}$) 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 enetered 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 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,[6] $P(k) \propto k^{n_s}/(1 + \alpha_p k + \beta_p k^2)^2$, with $\alpha_p = 8(\Omega_m h^2)^{-1}$ Mpc and $\beta_p = 4.7(\Omega_m h^2)^{-2}$ Mpc^2, with refinements that depend on the baryon mass fraction and neutrino properties (mass and number of flavors).[7] 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

[ii] 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.

of galaxies, clusters of galaxies, or the intergalactic gas. Computer codes that provide the detailed shape of the power-spectrum are available at **http://camb.info/** and **http://www.cmbfast.org**

Species that decouple 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. The 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 decouples from the cosmic plasma.

In order to determine the formation of objects of a given size or mass it is useful to consider the statistical distribution of the smoothed density field. To smooth the density distribution, cosmologists use a window (or filter) function $W(\mathbf{r})$ normalized so that $\int d^3r\, W(\mathbf{r}) = 1$, with the smoothed density perturbation field being $\int d^3r\delta(\mathbf{x})W(\mathbf{r})$. For the particular choice of a spherical top-hat window (similar to a cookie cutter), in which $W = 1$ in a sphere of radius R and $W = 0$ outside the sphere, the smoothed perturbation field measures the fluctuations in the mass in spheres of radius R. The normalization of the present power spectrum at $z = 0$ is often specified by the value of $\sigma_8 \equiv \sigma(R = 8h^{-1}\mathrm{Mpc})$ where $h = 0.7$ calibrates the Hubble constant today as $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$. 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. The variance $\langle\delta_M^2\rangle$ is

$$\sigma^2(M) \equiv \sigma^2(R) = \int_0^\infty \frac{dk}{2\pi^2}\, k^2 P(k) \left[\frac{3j_1(kR)}{kR}\right]^2 , \tag{2.7}$$

where $j_1(x) = (\sin x - x\cos x)/x^2$. The term involving j_1 in the integrand is the Fourier transform of $W(\mathbf{r})$. The function $\sigma(M)$ plays a crucial role in estimates of the abundance of collapsed objects, and is plotted in Figure 2.1 as a function of mass and redshift for the standard cosmological model. For modes with random phases, the probability of different regions with the same size to have a perturbation amplitude between δ and $\delta + d\delta$ is Gaussian with a zero mean and the above variance, $P(\delta)d\delta = (2\pi\sigma^2)^{-1/2}\exp\{-\delta^2/2\sigma^2\}d\delta$.

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

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

$$T(\theta) = T_\gamma\left(1 + \frac{v}{c}\cos\theta\right), \tag{2.8}$$

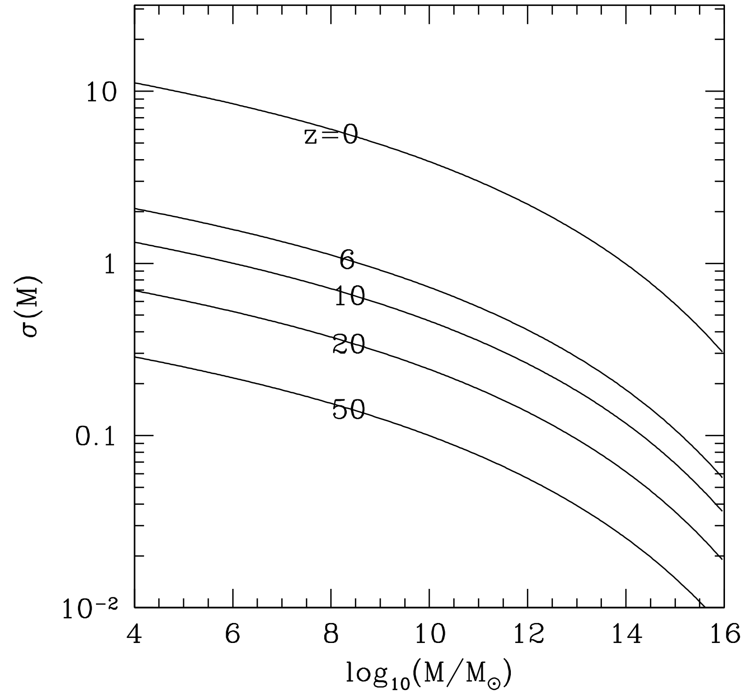

Figure 2.1 The root-mean-square amplitude of linearly-extrapolated density fluctuations σ as a function of mass M (in solar masses $M_\odot$, within a spherical top-hat filter) at different redshifts z. Halos form in regions that exceed the background density by a factor of order unity. This threshold 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 $R = 1.85\ \mathrm{Mpc}(M/10^{12}M_\odot)^{1/3}$.

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,

$$m_e \frac{dv}{dt} = -\int \sigma_T \cos\theta d\epsilon = -\frac{4}{3c}\sigma_T a T_\gamma^4 v. \tag{2.9}$$

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

$$\frac{d}{dt}E = -\frac{8\sigma_T}{3m_e c} a T_\gamma^4 E, \tag{2.10}$$

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,

$$\frac{dT_e}{dt} = \frac{x}{(1+x)} \frac{8\sigma_T a T^4}{3m_e c} (T_\gamma - T_e) - 2HT_e, \tag{2.11}$$

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 (2.11), $-2HT_e$, yields the adiabatic scaling $T_e \propto (1+z)^2$ in the absence of energy exchange with the CMB. More generally, this second term can be written as $(\gamma - 1)(\dot{\rho}_b/\rho_b)T_e$, where the ρ_b is the baryon density and $\gamma = \frac{5}{3}$ is the adiabatic index of a monoatomic gas. For the Hubble expansion, $(\dot{\rho}/\rho) = -3H$, while in overdense region, where expansion is replaced by contraction, this term changes sign and results in adiabatic heating (whereas the Compton cooling term remains unchanged).

The residual fraction of free electrons after cosmological recombination keeps the cosmic gas in thermal equilibrium with the CMB down to a redshift $z_t \sim 160$. Following reionization, 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 $z \sim 6$.

Chapter Three

Nonlinear Structure

3.1 COSMOLOGICAL JEANS MASS

As the density contrast between a spherical gas cloud and its cosmic environment grows, there are two main forces which come into play. The first is **gravity** and the second is **pressure**. We can get 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.[i] If the sound wave does not have sufficient time to traverse the cloud during the free-fall time, namely $R > R_J \equiv c_s t_{\rm coll}$, 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 a sufficiently small cloud with $R < R_J$. The transition between these regimes is defined by the so-called Jeans radius, R_J, corresponding to the Jeans mass,

$$M_J = \frac{4\pi}{3}\langle\rho\rangle R_J^3. \tag{3.1}$$

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 $M_J \sim 10^5 M_\odot$ is independent of redshift.[8] This is the minimum total mass of the first gas cloud 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.[9] Below we derive the above estimates more rigorously in the cosmological context of an expanding Universe.

[i] 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).

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,

$$\frac{\rho_{\rm dm}(R,t) - \bar{\rho}_{\rm dm}(t)}{\bar{\rho}_{\rm dm}(t)} = \delta_{\rm dm}(t)\frac{\sin(kR)}{kR} , \tag{3.2}$$

$$\frac{\rho_{\rm b}(R,t) - \bar{\rho}_{\rm b}(t)}{\bar{\rho}_{\rm b}(t)} = \delta_{\rm b}(t)\frac{\sin(kR)}{kR} , \tag{3.3}$$

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 a 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 evolution of the temperature of the baryons $T_{\rm b}(R, t)$ in the linear regime is determined by the coupling of their free electrons to the CMB through Compton scattering, and by the adiabatic expansion of the gas. Hence, $T_{\rm b}(r, t)$ is generally somewhere between the CMB temperature, $T_\gamma \propto (1 + z)^{-1}$ and the adiabatically-scaled temperature $T_{\rm ad} \propto (1 + z)^{-2}$. In the limit of tight coupling to T_γ, the gas temperature remains uniform. On the other hand, in the adiabatic limit, the temperature develops a gradient according to the relation

$$T_{\rm b} \propto \rho_b^{(\gamma-1)}. \tag{3.4}$$

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

$$\ddot{\delta}_{\rm dm} + 2H\dot{\delta}_{\rm dm} = \frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) \tag{3.5}$$

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

$$\begin{aligned}\ddot{\delta}_{\rm b} + 2H\dot{\delta}_{\rm b} = &\frac{3}{2}H^2\left(\Omega_{\rm b}\delta_{\rm b} + \Omega_{\rm dm}\delta_{\rm dm}\right) - \\ &\frac{kT_{\rm i}}{\mu m_p}\left(\frac{k}{a}\right)^2\left(\frac{a_{\rm i}}{a}\right)^{(1+\beta_T)}\left(\delta_{\rm b} + \frac{2}{3}\beta_T[\delta_{\rm b} - \delta_{\rm b,i}]\right).\end{aligned} \tag{3.6}$$

Here, $H(t) = \dot{a}/a$ is the Hubble parameter at a cosmological time t, and $\mu = 1.22$ is the mean atomic weight of the neutral primordial gas in units of the proton mass. The parameter β_T distinguishes between the two limits for the evolution of the gas temperature. In the adiabatic limit $\beta_T = 1$, and when the baryon temperature is uniform and locked to the background radiation, $\beta_T = 0$. The last term on the right hand side (in square brackets) takes into account the extra pressure gradient force in $\nabla(\rho_{\rm b}T) = (T\nabla\rho_{\rm b} + \rho_{\rm b}\nabla T)$, arising from the temperature gradient which develops 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 (3.6) to zero, and solving for the critical wavenumber $k_{\rm J}$. As can be seen from equation (3.6), the critical wavelength $\lambda_{\rm J}$ (and therefore the mass $M_{\rm J}$) is in general time-dependent. We infer from equation (3.6) that as time proceeds, perturbations with increasingly smaller initial wavelengths stop oscillating and start to grow.

To estimate the Jeans wavelength, we equate the right-hand-side of equation (3.6) to zero. We further approximate $\delta_{\rm b} \sim \delta_{\rm dm}$, and consider sufficiently high redshifts at which the Universe is matter dominated, $(1+z) \gg (\Omega_\Lambda/\Omega_m)^{1/3}]$. In this regime, $\Omega_{\rm b} \ll \Omega_m \approx 1$, $H \approx 2/(3t)$, and $a = (1+z)^{-1} \approx (3H_0\sqrt{\Omega_m}/2)^{2/3}t^{2/3}$, where $\Omega_m = \Omega_{\rm dm} + \Omega_b$ is the total matter density parameter. 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

$$(1+z_t) \approx 160(\Omega_b h^2/0.022)^{2/5}. \tag{3.7}$$

In the redshift range between recombination and z_t, $\beta_T = 0$ and

$$k_{\rm J} \equiv (2\pi/\lambda_{\rm J}) = [2kT_\gamma(0)/3\mu m_p]^{-1/2}\sqrt{\Omega_m}H_0\ , \tag{3.8}$$

so that the Jeans mass is redshift independent and obtains a value (for the total mass of baryons and dark matter),

$$M_{\rm J} \equiv \frac{4\pi}{3}\left(\frac{\lambda_{\rm J}}{2}\right)^3\bar{\rho}(0) = 1.35\times10^5\left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} M_\odot\ . \tag{3.9}$$

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,

$$M_{\rm J} = 4.54\times10^3\left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2}\left(\frac{\Omega_b h^2}{0.022}\right)^{-3/5}\left(\frac{1+z}{10}\right)^{3/2} M_\odot. \tag{3.10}$$

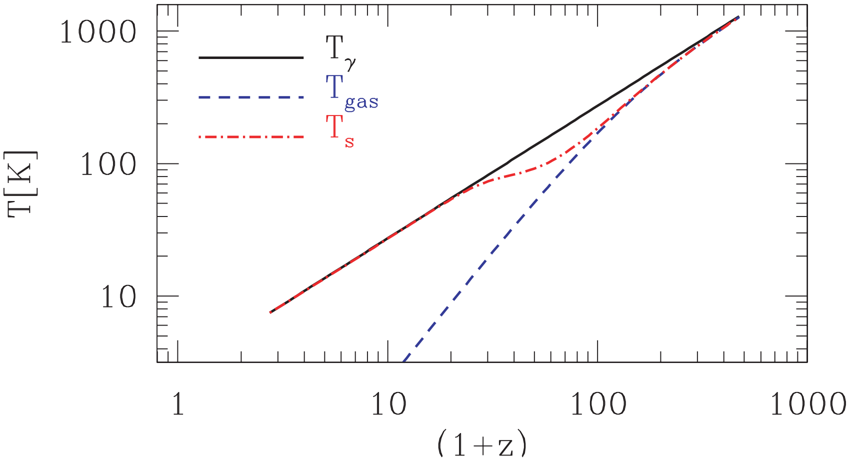

Figure 3.1 Thermal history of the baryons, left over from the big bang, before the first galaxies formed. The residual fraction of free electrons couple the gas temperture $T_{\rm gas}$ to the cosmic microwave background temperature $[T_\gamma \propto (1+z)]$ until a redshift $z \sim 200$. Subsequently the gas temperature cools adiabatically at a faster rate $[T_{\rm gas} \propto (1+z)^2]$. Also shown is the spin temperature of the 21cm transition of hydrogen $T_{\rm s}$ which interpolates between the gas and radiation temperature and will be discussed in Chapter 10.

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 3.5 and 3.6 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 correct time-weighted mass is the filtering mass[10] $M_F = (4\pi/3)\,\bar{\rho}\,(2\pi a/k_F)^3$, in terms of the comoving wavenumber k_F associated with the "filtering scale" (note the change in convention from π/k_J to $2\pi/k_F$). The wavenumber k_F is related to the Jeans wavenumber $k_{\rm J}$ by

$$\frac{1}{k_F^2(t)} = \frac{1}{D(t)} \int_0^t dt'\, a^2(t') \frac{\ddot{D}(t') + 2H(t')\dot{D}(t')}{k_J^2(t')} \int_{t'}^t \frac{dt''}{a^2(t'')}\,, \tag{3.11}$$

where $D(t)$ is the linear growth factor. At high redshift (where $\Omega_m(z) \to 1$), this relation simplifies to

$$\frac{1}{k_F^2(t)} = \frac{3}{a} \int_0^a \frac{da'}{k_J^2(a')} \left(1 - \sqrt{\frac{a'}{a}}\right)\ . \tag{3.12}$$

Then the relationship between the linear overdensity of the dark matter $\delta_{\rm dm}$ and the linear overdensity of the baryons δ_b, in the limit of small k, can be written as

$$\frac{\delta_b}{\delta_{\rm dm}} = 1 - \frac{k^2}{k_F^2} + O(k^4) \ . \tag{3.13}$$

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 nonlinear baryonic objects resulting from the shell-crossing and virialization of the dark matter, we must use a different model which examines the response of the gas to the gravitational potential of a virialized dark matter halo.

3.1.1 Spherical Collapse

Existing cosmological data suggests that the dark matter is "cold," that is, its pressure is negligible during the gravitational growth of galaxies. In popular models, 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.[11] All halos between this minimum clump mass and $\sim 10^5 M_\odot$ are expected to contain mostly dark matter and little ordinary matter. In describing the synamics of dark matter particles on large scales, we may ignore pressure and consider only the gravitational force.

For simplicity, let us consider a 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 reduce 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 or peculiar 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 profiles.

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

$$\frac{d^2r}{dt^2} = H_0^2 \Omega_\Lambda \, r - \frac{GM}{r^2} \ , \tag{3.14}$$

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 (3.14) as

$$\frac{l}{H_0^2}\frac{d^2x}{dt^2} = -\frac{\Omega_m}{2x^2}(1+\delta_i) + \Omega_\Lambda x, \tag{3.15}$$

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

$$x(t_i) = a_i \tag{3.16}$$

$$\frac{dx}{dt}(t_i) = H(t_i)x\left(1 - \frac{\delta_i}{3}\right) = H_0 a_i\left(1 - \frac{\delta_i}{3}\right)\sqrt{\frac{\Omega_m}{a_i^3} + \Omega_\Lambda}, \tag{3.17}$$

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 (3.15) yields

$$\frac{1}{H_0^2}\left(\frac{dx}{dt}\right)^2 = \frac{\Omega_m}{x}(1+\delta_i) + \Omega_\Lambda x^2 + K, \tag{3.18}$$

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

$$K = -\frac{5\delta_i}{3a_i}\Omega_m. \tag{3.19}$$

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

$$H_0 t_{coll} = 2\int_0^{a_{\max}} da\left(\Omega_m/a + K + \Omega_\Lambda a^2\right)^{-1/2}, \tag{3.20}$$

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:

$$\left.\begin{aligned} r &= A(\cos\eta - 1) \\ t &= B(\eta - \sin\ \eta) \end{aligned}\right\} \qquad \text{Closed} \qquad (0 \le \eta \le 2\pi) \tag{3.21}$$

$$\left.\begin{aligned} r &= A\eta^2/2 \\ t &= B\eta^3/6 \end{aligned}\right\} \qquad \text{Flat} \qquad (0 \le \eta \le \infty) \tag{3.22}$$

$$\left.\begin{aligned} r &= A(\cosh\eta - 1) \\ t &= B(\sinh\eta - \eta) \end{aligned}\right\} \qquad \text{Open} \qquad (0 \le \eta \le \infty) \tag{3.23}$$

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 $2\pi B$.

We are now faced with the problem of relating the spherical collapse parameters $A, B,$ and M to the linear theory density perturbation δ. For the case of Ω_Λ = 0 and $\Omega_m = 1$, we can determine the spherical collapse parameters A and B. $K > 0$ $(K < 0)$ produces an open (closed) model. Comparing coefficients in the energy equation (3.18) and the integral of the equation of motion, one finds

$$A = \frac{r_i}{2a_i}\left(\frac{5\delta_i}{3a_i}\right)^{-1} \tag{3.24}$$

$$B = \frac{1}{2H_0}\left(\frac{5\delta_i}{3a_i}\right)^{-3/2} . \tag{3.25}$$

In an $\Omega = 1$ Universe, where $1 + z = (3H_0 t/2)^{-2/3}$, we find that a shell collapses at redshift $1 + z_c = 0.5929\delta_i/a_i$, or in other words a shell collapsing at redshift z_c had a linear overdensity extrapolated to the present day[ii] of $\delta_0 = 1.686(1 + z_c)$.

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 $\overline{\delta_i}$

$$\overline{\delta_i}(R) = \frac{3}{4\pi R^3}\int_0^R d^3r\delta_i(r), \tag{3.26}$$

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 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

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

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[12] (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 (1.4) for a spherical overdense region, it is possible to relate the characteristic radius and gravitational potential well of each of these galaxies to their mass and their redshift of formation.

The small density fluctuations evidenced in the CMB grew over time as described in §2.1, until the perturbations δ became of order unity and the full non-linear gravitational collapse followed. The dynamical collapse of a dark matter halo can be solved analytically in spherical symmetry with an initial top-hat of uniform overdensity δ_i inside a sphere of radius R. Although this toy model might seem artificially simple, its results have turned out to be surprisingly accurate for interpreting the properties and distribution of halos in numerical simulations of cold dark matter.

During the gravitational collapse of a spherical region, the enclosed overdensity δ grows initially as $\delta_L = \delta_i D(t)/D(t_i)$, in accordance with linear theory, but eventually δ grows above δ_L. Any mass shell that is gravitationally bound (i.e., with a negative total Newtonian energy) reaches a radius of maximum expansion (turn-around) and subsequently collapses. The solution of the equation of motion for a top-hat region shows that at the moment when the region collapses to a point, the overdensity predicted by linear theory is $\delta_L = 1.686$ in the $\Omega_m = 1$ case, with only a weak dependence on Ω_Λ in the more general case. Thus, a top-hat would have collapsed at redshift z if its linear overdensity extrapolated to the present day (also termed the critical density of collapse) is

$$\delta_{\rm crit}(z) = \frac{1.686}{D(z)} \, , \tag{3.27}$$

where we set $D(z = 0) = 1$.

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,[iii] which applies at redshifts $z \gg 1$. We restrict our attention below to these high redshifts.

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

$$r_{\rm vir} = 1.5 \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{-1} \ {\rm kpc}\ , \tag{3.28}$$

and a corresponding circular velocity,

$$V_c = \left(\frac{GM}{r_{\rm vir}}\right)^{1/2} = 17.0 \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{1/2} \ {\rm km\ s^{-1}}\ . \tag{3.29}$$

We may also define a virial temperature

$$T_{\rm vir} = \frac{\mu m_p V_c^2}{2k} = 1.04 \times 10^4 \ \left(\frac{\mu}{0.6}\right) \left(\frac{M}{10^8 M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \ {\rm K}\ , \tag{3.30}$$

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,

$$E_b = \frac{1}{2}\frac{GM^2}{r_{\rm vir}} = 2.9 \times 10^{53} \left(\frac{M}{10^8 M_\odot}\right)^{5/3} \left(\frac{1+z}{10}\right) \ {\rm erg}\ . \tag{3.31}$$

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[iv] ($\sim 10^{51}$ 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 profile has the form[v]

$$\rho(r) = \frac{3H_0^2}{8\pi G}(1+z)^3\Omega_m \frac{\delta_c}{c_{\rm N}x(1+c_{\rm N}x)^2}\ , \tag{3.32}$$

[iii] This implies that 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.

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

[v] 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).

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

$$\delta_c = \frac{\Delta_c}{3} \frac{c_{\rm N}^3}{\ln(1+c_{\rm N}) - c_{\rm N}/(1+c_{\rm N})} \ . \tag{3.33}$$

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.

3.3 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.[13] 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 a 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. Since δ_M is distributed as a Gaussian variable with zero mean and standard deviation $\sigma(M)$ (which depends only on the present linear power spectrum; see equation 2.7), the probability that δ_M is greater than some δ equals

$$\int_\delta^\infty d\delta_M \frac{1}{\sqrt{2\pi}\,\sigma(M)} \exp\left[-\frac{\delta_M^2}{2\,\sigma^2(M)}\right] = \frac{1}{2}{\rm erfc}\left(\frac{\delta}{\sqrt{2}\,\sigma(M)}\right) \ . \tag{3.34}$$

The basic ansatz is to identify this probability with the fraction of dark matter particles which are part of collapsed halos of mass greater than M at redshift z. There are two additional ingredients. First, the value used for δ is $\delta_{\rm crit}(z)$ (given in equation 3.27), which is the critical density of collapse found for a spherical top-hat (extrapolated to the present since $\sigma(M)$ is calculated using the present power spectrum at $z = 0$); and second, the fraction of dark matter in halos above M is multiplied by an additional factor of 2 in order to ensure that every particle ends up as part of some halo with $M > 0$. Thus, the final formula for the mass fraction in halos above M at redshift z is

$$F(>M|z) = {\rm erfc}\left(\frac{\delta_{\rm crit}(z)}{\sqrt{2}\,\sigma(M)}\right) \ . \tag{3.35}$$

Differentiating the fraction of dark matter in halos above M yields the mass distribution. Letting dn be the comoving number density of halos of mass between

M and $M + dM$, we have

$$\frac{dn}{dM} = \sqrt{\frac{2}{\pi}}\,\frac{\rho_m}{M}\,\frac{-d(\ln\sigma)}{dM}\,\nu_c\,e^{-\nu_c^2/2}\ , \tag{3.36}$$

where $\nu_c = \delta_{\rm crit}(z)/\sigma(M)$ is the number of standard deviations which 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.

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 halos at high redshift. The halo mass function of Sheth & Tormen (1999) adds two free parameters that allow it to fit numerical simulations much more accurately,

$$\frac{dn}{dM} = A'\sqrt{\frac{2a'}{\pi}}\,\frac{\rho_m}{M}\,\frac{-d(\ln\sigma)}{dM}\,\nu_c\,\left[1+\frac{1}{(a'\nu_c^2)^{q'}}\right]e^{-a'\nu_c^2/2}\ , \tag{3.37}$$

with best-fit parameters $a' = 0.75$ and $q' = 0.3$, and where proper normalization is ensured by adopting $A' = 0.322$. Results for the associated comoving density of halos of different masses at different redshifts are shown in Figure 3.2.

The ad-hoc factor of 2 in the Press-Schechter derivation is necessary, since otherwise only positive fluctuations of δ_M would be included. Bond et al. (1991) found an alternate derivation of this correction factor, using a different ansatz, called the excursion set (or extended Press-Schechter) formalism.[14] In their derivation, the factor of 2 has a more satisfactory origin. For a given mass M, even if δ_M is smaller than $\delta_{\rm crit}(z)$, it is possible that the corresponding region lies inside a region of some larger mass $M_L > M$, with $\delta_{M_L} > \delta_{\rm crit}(z)$. In this case the original region should be counted as belonging to a halo of mass M_L. Thus, the fraction of particles which are part of collapsed halos of mass greater than M is larger than the expression given in equation (3.34).

3.3.1 The Excursion-Set (Extended Press-Schechter) Formalism

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 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 $\overline{\delta}(R)$, the average overdensity within some spherical window of characteristic radius R, as a function of R. While the Press-Schechter model depends only on the Gaussian distribution of $\overline{\delta}$ for one particular R, the excursion set considers all R. Again the connection between a value of the linear

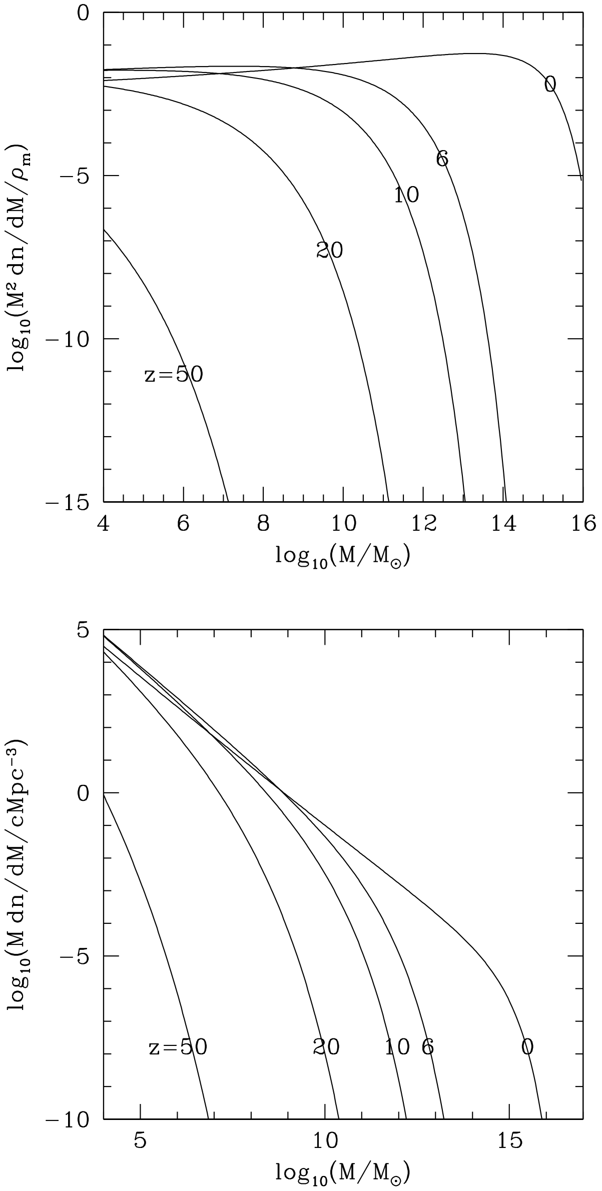

Figure 3.2 *Top:* The mass fraction incorporated into halos per logarithmic bin of halo mass $(M^2dn/dM)/\rho_m$, as a function of M at different redshifts z. Here $\rho_m = \Omega_m\rho_c$ is the present-day matter density, and $n(M)dM$ is the comoving density of halos with masses between M and $M+dM$. 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.

regime δ and the final state is made via the spherical collapse solution so that there is a critical value $\delta_{\rm crit}(z)$ of $\overline{\delta}$ which is required for collapse at a redshift z.

For most choices of window function, the functions $\overline{\delta}(R)$ are correlated from one R 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 R greatly simplify and many interesting quantities may be calculated.[15] The key is to 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 $k > k_c$. This filter has a spatial form of $W(r) \propto j_1(k_c r)/k_c r$, which 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. Note that in real space, this window function exhibits a sinusoidal oscillation and is not sharply localized.

The great advantage of the sharp k-space filter is that the difference at a given point between $\overline{\delta}$ on one mass scale and that on another mass scale is statistically independent from the value on the larger mass scale. With a Gaussian random field, each δ_k is Gaussian distributed independently from the others. For this filter,

$$\overline{\delta}(M) = \int_{k<k_c(M)} \frac{d^3k}{(2\pi)^3}\delta_k, \tag{3.38}$$

meaning that the overdensity on a particular scale is simply the sum of the random variables δ_k interior to the chosen k_c. Consequently, the difference between the $\overline{\delta}(M)$ 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 from the sum of the δ_k interior to the smaller k_c. Meanwhile, the distribution of $\overline{\delta}(M)$ given no prior information is still a Gaussian of mean zero and variance

$$\sigma^2(M) = \frac{1}{2\pi^2}\int_{k<k_c(M)} dk\, k^2 P(k). \tag{3.39}$$

If we now consider $\overline{\delta}$ as a function of scale k_c, we see that we begin from $\overline{\delta} = 0$ at $k_c = 0$ $(M = \infty)$ and then add independently random pieces as k_c increases. This generates a random walk, albeit one whose stepsize varies with k_c. We then assume that at redshift z, a given function $\overline{\delta}(k_c)$ represents a collapsed mass M corresponding to the k_c where the function first crosses the critical value $\delta_{\rm crit}(z)$. With this assumption, we may use the properties of random walks to calculate the evolution of the mass as a function of redshift.

It is now easy to re-derive the Press-Schechter mass function, including the previously unexplained factor of 2. 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)* $\overline{\delta}(K_c) > \delta_{\rm crit}(z)$, or that *(b)* $\overline{\delta}(K_c) < \delta_{\rm crit}(z)$ but $\overline{\delta}(k_c') > \delta_{\rm crit}(z)$ for some $k_c' < K_c$. But these two cases in fact have 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. Since the distribution of $\overline{\delta}(K_c)$ is simply Gaussian with variance $\sigma^2(M)$, the fraction of random walks falling into class *(a)* is simply $(1/\sqrt{2\pi\sigma^2})\int_{\delta_{\rm crit}(z)}^{\infty} d\delta\ \exp\{-\delta^2/2\sigma^2(M)\}$. Hence, the fraction of mass elements included in halos of mass less than M at redshift z is simply

$$F(< M) = 1 - 2 \times \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\delta_{\rm crit}(z)}^{\infty} d\delta \, \exp\left\{ -\frac{\delta^2}{2\sigma^2(M)} \right\}, \tag{3.40}$$

which may be differentiated to yield the Press-Schechter mass function. We may now go further and consider how halos at one redshift are related to those at another redshift. If it is given that a halo of mass M_2 exists at redshift z_2, then we know that the random function $\overline{\delta}(k_c)$ for each mass element within the halo first crosses $\delta(z_2)$ at k_{c2} corresponding to M_2. Given this constraint, we may study the distribution of k_c where the function $\overline{\delta}(k_c)$ crosses other thresholds. 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 $z_1 > z_2$); clearly this occurs at some $k_{c1} > k_{c2}$. This problem reduces to the previous one if we translate the origin of the random walks from $(k_c, \overline{\delta}) = (0, 0)$ to $(k_{c2}, \delta_{\rm crit}(z_2))$. We therefore find 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, is

$$\frac{dP}{dM_1}(M_1, z_1|M_2, z_2) = \\ \sqrt{\frac{2}{\pi}} \frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{[\sigma^2(M_1) - \sigma^2(M_2)]^{3/2}} \left| \frac{d\sigma(M_1)}{dM_1} \right| \exp\left\{ -\frac{[\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)]^2}{2[\sigma^2(M_1) - \sigma^2(M_2)]} \right\}. \tag{3.41}$$

This may be rewritten as saying that the quantity

$$\tilde{v} = \frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{\sqrt{\sigma^2(M_1) - \sigma^2(M_2)}} \tag{3.42}$$

is distributed as the positive half of a Gaussian with unit variance; equation (3.42) 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 look back in time, the object breaks into ever smaller pieces, similar to the branching of a tree. Equation (3.41) is the distribution of the sizes of these branches at some given earlier time. However, using this description of the ensemble distribution to generate random realizations of single merger trees has proven to be difficult. In all cases, one recursively steps back in time, at each step breaking the final object into two or more pieces. A simplified scheme may assume that at each time step, the object breaks into only two pieces. One value from the distribution (3.41) then determines the mass ratio of the two branches.

We may also use the distribution of the ensemble to derive some additional analytic results. 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:

$$P(z_f > z_1) = \int_{M_2/2}^{M_2} \frac{M_2}{M} \frac{dP}{dM}(M, z_1|M_2, z_2) dM, \tag{3.43}$$

where dP/dM is given in equation (3.41). The factor of M_2/M corrects the counting from being mass weighted to number weighted; each halo of mass M_2 can have only one progenitor of mass greater than $M_2/2$. Differentiating equation (3.43) with respect to time gives the distribution of formation times. Overall, the excursion set formalism provides a good approximation to more exact numerical simulations of halo assembly and merging histories.

3.4 NONLINEAR CLUSTERING: THE HALO MODEL

3.5 NUMERICAL SIMULATIONS OF STRUCTURE FORMATION

Chapter Four

The Intergalactic Medium

4.1 THE LYMAN-α FOREST

4.1.1 The Ionizing Background

4.2 METAL-LINE SYSTEMS

4.3 THEORETICAL MODELS

4.3.1 Numerical Simulations

4.3.2 Semi-Analytic Models

Chapter Five

The First Stars

5.1 CHEMISTRY AND COOLING OF PRIMORDIAL GAS

When a dark matter halo collapses, the associated gas falls in at a speed comparable to V_c in equation (3.29). When multiple gas streams collide and settle to a static configuration, the gas shocks to the virial temperature $T_{\rm vir}$ in equation (3.30) – 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 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),[i] to which the cosmic gas is transparent even before it is ionized.[16] 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 $T_{\rm vir}$.

As soon as halos with $T_{\rm vir} > 10^4$K formed, atomic hydrogen was able to cool the gas in them and allow fragmentation even in the absence of H_2. In addition, once

[i] 1 electron Volt (eV) is an energy unit equivalent to 1.6×10^{-12}ergs or 11,604K.

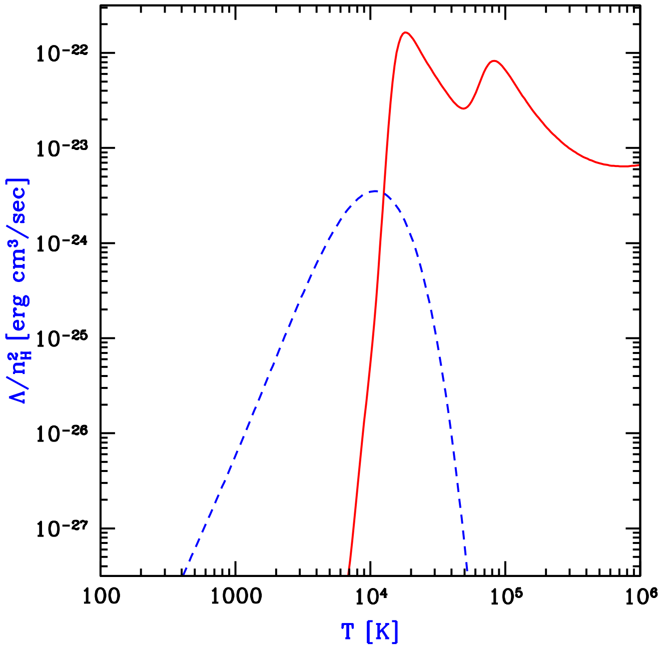

Figure 5.1 Cooling rates as a function of temperature for a primordial gas composed of atomic hydrogen and helium, as well as molecular hydrogen, in the absence of any external radiation. We assume a hydrogen number density $n_H = 0.045\ \mathrm{cm}^{-3}$, corresponding to the mean density of virialized 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.

the gas was enriched with heavy elements, it was able to cool even more efficiently.

5.1.1 Chemistry

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

$$\begin{aligned} \mathrm{H} \;\; &+ \;\; \mathrm{e}^- \;\rightarrow\; \mathrm{H}^- \;\; + \;\; h\nu, && (5.1)\\ \mathrm{H}^- \;\; &+ \;\; \mathrm{H} \;\rightarrow\; \mathrm{H}_2 \;\; + \;\; \mathrm{e}^-, && (5.2) \end{aligned}$$

where free electrons act as catalysts. 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. Table 5.2 shows the same for deuterium mediated reactions. Due to the low gas density, the chemical reactions are slow and the molecular abundance is far from its value in chemical equilibrium. After cosmological recombination and before the first galaxies had formed, the fractional H_2 abundance is very small ($\sim 6 \times 10^{-7}$) relative to hydrogen by number.[17] At redshifts $z \ll 100$, the gas temperature in most regions is too low for collisional ionization to be effective, and free electrons (over and above the residual electron fraction) are mostly produced through photoionization of neutral hydrogen by UV or X-ray radiation from stars.

In objects with baryonic masses $> 3 \times 10^4 M_\odot$, gravity dominates and results in the bottom-up hierarchy of structure formation characteristic of CDM cosmologies; at lower masses, gas pressure delays the collapse. The first objects to collapse are those at the mass scale that separates these two regimes. Such objects reach virial temperatures of several hundred degrees and can fragment into stars only through cooling by molecular hydrogen. In other words, there are two independent minimum mass thresholds for star formation: the Jeans mass (related to accretion) 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 $z \sim 20$.

However, molecular hydrogen (H_2) is fragile and can easily be photo-dissociated by photons with energies of 11.26–13.6eV, to which the IGM is transparent even before it is ionized. The photo-dissociation occurs through a two-step process, first suggested by Phil Solomon in 1965 and later analyzed quantitatively[18] by Stecher & Williams (1967). Haiman, Rees, & Loeb (1997) evaluated the average cross-section for this process between 11.26eV and 13.6eV, by summing the oscillator strengths for the Lyman and Werner bands of H_2, and obtained a value of $3.71 \times 10^{-18}\ \mathrm{cm}^2$. They showed that the UV flux capable of dissociating H_2 throughout the collapsed environments in the universe is lower by more than two orders of magnitude than the minimum flux necessary to ionize the universe. The inevitable conclusion is that soon after trace amounts of stars form, the formation of additional stars due to H_2 cooling is suppressed. Further fragmentation is possible only through atomic line cooling, which is effective in objects with much higher virial temperatures, $T_{\mathrm{vir}} > 10^4$K. Such objects correspond to a total mass $> 10^8 M_\odot[(1+z)/10]^{-3/2}$. Figure 5.2 illustrates this sequence of events by describing two classes of objects:

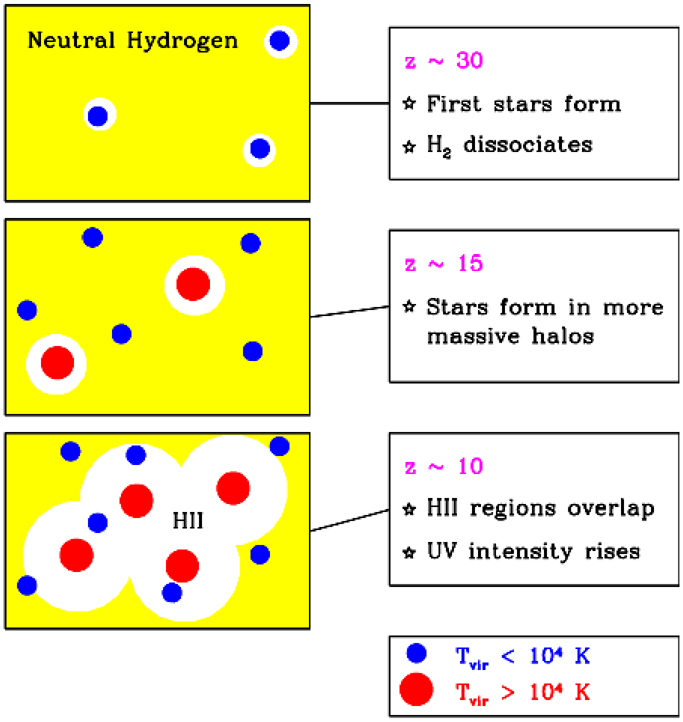

Figure 5.2 Stages in the reionization of hydrogen in the intergalactic medium.

those with $T_{\rm vir} < 10^4$K (small dots) and those with $T_{\rm vir} > 10^4$K (large dots). In the first stage (top panel), some low-mass objects collapse, form stars, and create ionized hydrogen (H II) bubbles around them. Once the UV background between 11.2–13.6eV reaches a specific critical level, H_2 is photo-dissociated throughout the universe and the formation of new stars is delayed until objects with $T_{\rm vir} > 10^4$K collapse.

When considering the photo-dissociation of H_2 before reionization, it is important to incorporate the *processed* spectrum of the UV background at photon energies below the Lyman limit. Due to the absorption at the Lyman-series resonances this spectrum obtains the sawtooth shape shown in Figure 5.3. 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 Lyman-series resonances along the way. 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 $z = 11.1$.

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)	$\mathrm{H} + e^- \to \mathrm{H}^+ + 2e^-$	$5.85 \times 10^{-11} T^{1/2} \mathrm{exp}(-157,809.1/T)(1+T_5^{1/2})^{-1}$
(2)	$\mathrm{H}^+ + e^- \to \mathrm{H} + h\nu$	$8.40 \times 10^{-11} T^{-1/2} T_3^{-0.2}(1+T_6^{0.7})^{-1}$
(3)	$\mathrm{H} + e^- \to \mathrm{H}^- + h\nu$	$1.65 \times 10^{-18} T_4^{0.76+0.15\log_{10} T_4 - 0.033\log_{10}^2 T_4}$
(4)	$\mathrm{H} + \mathrm{H}^- \to \mathrm{H}_2 + e^-$	1.30×10^{-9}
(5)	$\mathrm{H}^- + \mathrm{H}^+ \to 2\mathrm{H}$	$7.00 \times 10^{-7} T^{-1/2}$
(6)	$\mathrm{H}_2 + e^- \to \mathrm{H} + \mathrm{H}^-$	$2.70 \times 10^{-8} T^{-3/2} \mathrm{exp}(-43,000/T)$
(7)	$\mathrm{H}_2 + \mathrm{H}^+ \to \mathrm{H}_2^+ + \mathrm{H}$	$2.40 \times 10^{-9} \mathrm{exp}(-21,200/T)$
(8)	$\mathrm{H}_2 + e^- \to 2\mathrm{H} + e^-$	$4.38 \times 10^{-10} \mathrm{exp}(-102,000/T) T^{0.35}$
(9)	$\mathrm{H}^- + e^- \to \mathrm{H} + 2e^-$	$4.00 \times 10^{-12} T \,\mathrm{exp}(-8750/T)$
(10)	$\mathrm{H}^- + \mathrm{H} \to 2\mathrm{H} + e^-$	$5.30 \times 10^{-20} T \,\mathrm{exp}(-8750/T)$

Thus, sources in the redshift interval 10–11.1 could be seen at 11 eV, but radiation emitted by sources 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 from the nearest Lyman resonance. Most of the photons absorbed along the way would be re-emitted at Lyman-α 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. 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.

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.

5.2 FORMATION OF THE FIRST METAL-FREE STARS

5.2.1 Sheets, Filaments, and Only Then, Galaxies

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

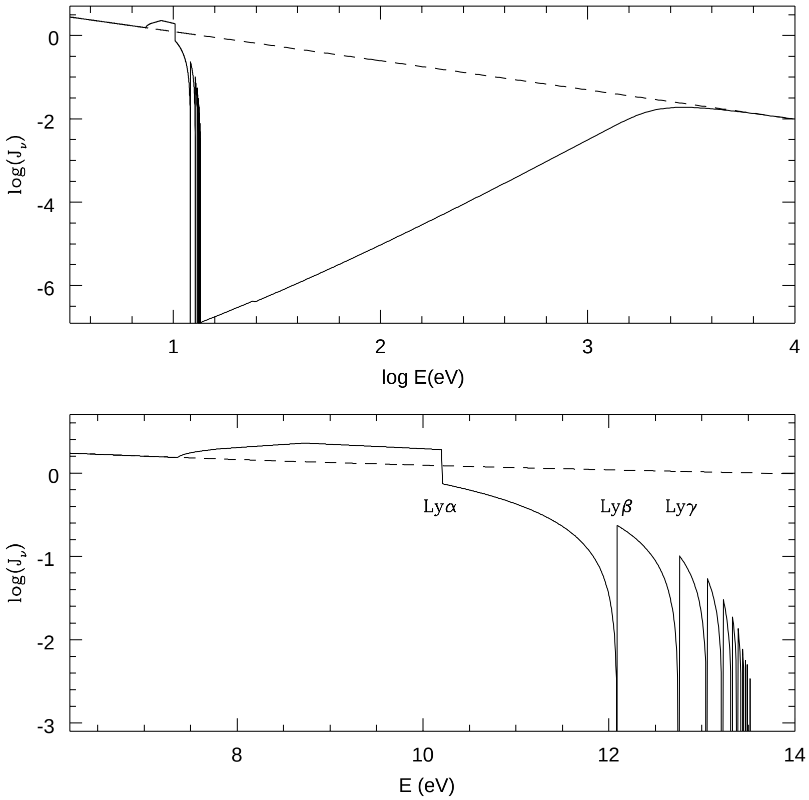

Figure 5.3 The average spectrum during the initial phase of the reionization epoch (arbitrary units). The upper panel shows that absorption by neutral hydrogen and helium suppresses the flux above 13.6eV up to the keV range. The lower panel shows a close-up of the sawtooth modulation due to line absorption below 13.6 eV. A constant comoving density of sources was assumed, with each source emitting a power-law continuum, which would result in the spectrum shown by the dashed lines if absorption were not taken into account. Figure credit: Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **476**, 458 (1997).

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$^+$	1.00×10^{-9}exp$(-464/T)$

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 5.4, 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 modelling 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.

Stars form in the densest, coolest knots of gas, in which the Jeans mass is lowered to the scale of a single star. By observing the radiation from galaxies, one is mapping the distribution of the densest peaks. The situation is analogous to a satellite image of the Earth at night in which light paints the special locations of big cities, while many other topographical details are hidden from view. It is, in principle, possible to probe the diffuse cosmic gas directly by observing its emission or absorption properties.

We will describe two methods for studying structures in the early Universe: *(i)* imaging the stars and black holes within the first galaxies, and *(ii)* imaging the diffuse gas in between these galaxies.

5.2.2 Metal-free Stars

The formation of the first stars hundreds of millions of years after the Big Bang marks the temporal boundary between these two branches. Earlier than that, 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. 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. Although the journey that led to our 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.

Gas cooling in nearby galaxies is affected mostly by heavy elements (in a variety of forms, including atoms, ions, molecules, and dust) which are produced in stellar interiors and get mixed into the interstellar gas by supernova explosions. These

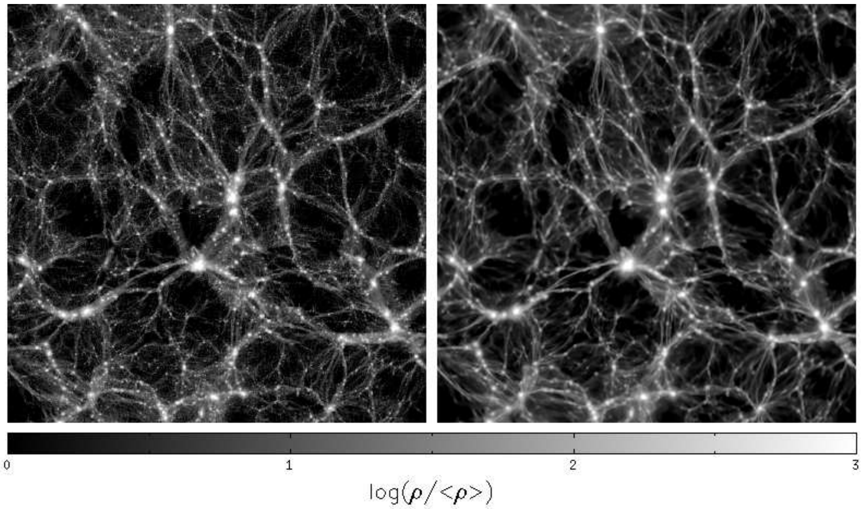

Figure 5.4 The large-scale distributions of dark matter (left) and gas (right) in the IGM show a network of filaments and sheets, known as the "cosmic web". Overall, the gas follows the dark matter on large scales but is more smoothly distributed on small scales owing to its pressure. The snapshots show the projected density contrast in a 7 Mpc thick slice at zero redshift from a numerical simulation of a box measuring 140 comoving Mpc on a side. Figure credit: Trac, H., & Pen, U.-L. *New Astron.* **9**, 443 (2004).

powerful explosions are triggered at the end of the life of massive stars after their core consumes its nuclear fuel reservoir, loses its pressure support against gravity, and eventually collapses to make a black hole or a compact star made of neutrons with the density of an atomic nucleus. A neutron star has a size of order ~ 10 kilometers – comparable to a big city – but contains a mass comparable to the Sun. As infalling material arrives at the surface of the proto-neutron star, it bounces back and sends a shock wave into the surrounding envelope of the star which then explodes, exporting heavy elements into the surrounding medium.

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.[19] 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. As a result, cooling of the primordial gas and its fragmentation into the first stars was initially mediated by trace amounts of molecular hydrogen in halos just above the cosmological Jeans mass of ~ 0.1–1 million solar masses ($T_{\rm vir} \sim$ hundreds of degrees K). Subsequently, star formation became much more efficient through the cooling of atomic hydrogen (see Figure 5.1) in the first dwarf galaxies that were at least a thousand times more massive ($T_{\rm vir} > 10^4$K; see equation 3.30). The evolution of star formation in the first galaxies was also shaped by a variety of feedback processes. Internal self-regulation involved feedback from vigorous episodes of star formation (through supernova-driven winds) and black hole accretion (through the intense radiation and outflows it generates). But there was also external feedback. The reionization of the intergalactic gas heated the gas and elevated its Jeans mass. After reionization the intergalactic gas could not have assembled into the shallowest potential wells of dwarf galaxies.[20] This suppression of gas accretion may explain the inferred deficiency of dwarf galaxies relative to the much larger population of dark matter halos that is predicted to exist by numerical simulations but not observed around the Milky Way.[21] If this interpretation of the deficiency is correct, then most of the low-mass halos that formed after reionization were left devoid of gas and stars and are therefore invisible today. But before we get to these late stages, let us start at the beginning and examine the formation sites of the very first stars.

How did the the first clouds of gas form and fragment into the first stars? This questions 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 simulate the collapse and fragmentation of the first gas clouds and the formation of stars within them.

Results from such numerical simulations of a collapsing halo with $\sim 10^6 M_\odot$ are presented in Figure 5.5. Generically, the collapsing region makes a central massive clump with a typical mass of hundreds of solar masses, which happens to be the Jeans mass for a temperature of ~ 500K and the density $\sim 10^4\ \mathrm{cm}^{-3}$ at which the gas lingers because its H_2 cooling time is longer than its collapse time at that point. Soon after its formation, the clump becomes gravitationally unstable and undergoes runaway collapse at a roughly constant temperature due to H_2 cooling. The central clump does not typically undergo further sub-fragmentation, and is expected to form a single star. Whether more than one star can form in a low-mass halo thus crucially depends on the degree of synchronization of clump formation,[22] since the radiation from the first star to form can influence the motion of the surrounding gas more than gravity.[23]

How massive were the first stars? 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. On dimensional grounds, the mass growth rate is simply given by the ratio between the Jeans mass and the free-fall time, implying $(dm_\star/dt) \sim c_s^3/G \propto T^{3/2}$. 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 rate of mass growth for the star typically tapers off with time.[24] A rough upper limit for the final mass of the star is obtained by continuing its accretion for its total lifetime of a few million years, yielding a final mass of $< 10^3 M_\odot$. *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

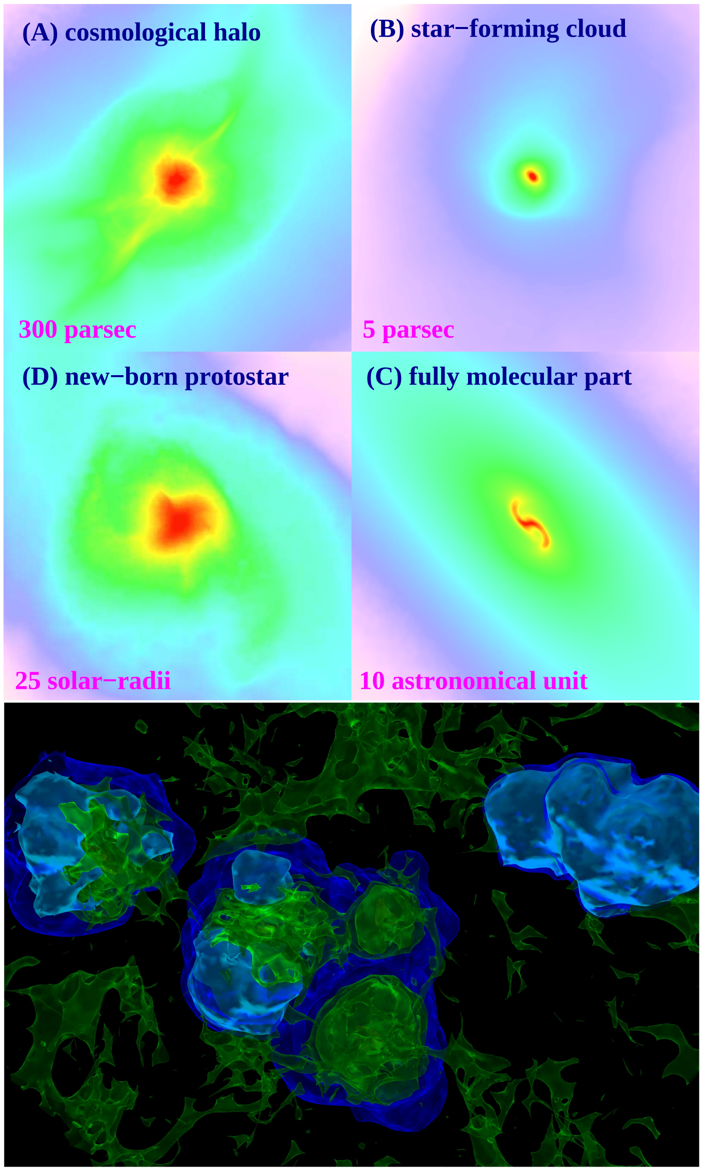

Figure 5.5 Results from a numerical simulation of the formation of a metal-free star [Yoshida, N., Omukai, K., & Hernquist, L. *Science* **321**, 669 (2008)] and its feedback on its environment [Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F. *Nature* **459**, 49 (2009)]. *Top:* Projected gas distribution around a primordial protostar. 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. *Bottom:* 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.

eventually curtailed by feedback from the star.

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.

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.[25] Population I 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 on large scales 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,[26] with a number of stars $N_\star$ per logarithmic bin of star mass $m_\star$,

$$\frac{dN_\star}{d\log m_\star} \propto {m_\star}^{-\Gamma}, \tag{5.3}$$

where

$$\Gamma \simeq \begin{cases} 1.35 & \text{for } m_\star > 0.5M_\odot \\ 0.0 & \text{for } 0.008M_\odot < m_\star < 0.5M_\odot \end{cases} . \tag{5.4}$$

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.[27] Moreover, nuclear fusion reactions do not ignite in the cores of proto-stars below a mass of $\sim 0.08M_\odot$, so-called "brown dwarfs". The most important feature of this IMF is that $\sim 1M_\odot$ characterizes the mass scale of Population I star formation, in the sense that most of the stellar mass goes into stars with masses close to this value.

Since current simulations indicate that the first stars were predominantly very massive ($> 30M_\odot$), and consequently rather different from present-day stellar populations, an interesting question arises: *how and when did the transition take place from the early formation of massive stars to the late-time formation of low-mass stars?*

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 these stars appeared, however, the situation became more complex due to their feedback on the environment. In particular, supernova explosions dispersed the heavy elements produced inside 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 was likely the dominant effect that brought about the transition from Population III to Population II star formation. Comparison of the cooling rate by singly-ionized carbon and neutral oxygen to that of H_2 indicates that as soon as the abundance of these elements exceeded a level as small as 0.1%

of the solar abundance (or even lower if dust forms), the cooling of the gas became much more efficient than that provided by H_2 molecules.[28] The characteristic mass scale for star formation is therefore expected to be a function of metallicity, with a sharp transition at this metallicity threshold, above which the characteristic mass of a star gets reduced by about two orders of magnitude. Nevertheless, one should keep in mind that the temperature floor of the gas was dictated by the CMB (whose temperature was $54.6 \times [(1+z)/20]$ K) and therefore even with efficient cooling, the stars at high redshifts were likely more massive than the stars found today.

The maximum distance out to which a galactic outflow mixes heavy elements with the IGM can be estimated based on energy considerations. The mechanical energy released E will accelerate all the gas it encounters into a thin shell at a physical distance $R_{\rm max}$ from the central source. In doing so, it must accelerate the swept-up gas to the Hubble flow velocity at that distance, $v_{\rm s} = H(z)R_{\rm max}$. If the shocked gas has a short cooling time, then its original kinetic energy is lost and is unavailable for expanding the shell. Ignoring the gravitational effect of the host galaxy, deviations from the Hubble flow, and cooling inside the cavity bounded by the shell, energy conservation implies: $E = \frac{1}{2}M_{\rm s}v_{\rm s}^2$, where $M_{\rm s} = \frac{4\pi}{3}\bar{\rho}R_{\rm max}^3$. At $z \gg 1$, this gives a maximum outflow distance: $R_{\rm max} \sim 50~{\rm kpc}(E/10^{56}~{\rm ergs})^{1/5}[(1+z)/10]^{-6/5}$. The maximum radius of influence from a galactic outflow can therefore be estimated based on the total number of supernovae that power it (each releasing $\sim 10^{51}~{\rm ergs}$ of which a substantial fraction may be lost by early cooling) or the mass $M_{\rm bh}$ of the central black hole (typically releasing a fraction of a percent of $M_{\rm bh}c^2$ in mechanical energy). More detailed calculations give similar results.[29] Finally, the fraction of the IGM enriched with heavy elements can be obtained by multiplying the density of galactic halos with their individual volumes of influence.

Since the earliest galaxies represent high density peaks and are therefore clustered, the metal enrichment process was inherently non-uniform. The early evolution of the volume filling of metals in the IGM can be inferred from the spectra of bright high-redshift sources.[30] Even at late cosmic times, it should be possible to find regions of the Universe that are composed of primordial gas and hence could make Population III stars. Since massive stars produce ionizing photons much more effectively than low-mass stars, the transition from Population III to Population II stars had important consequences for the ionization history of the cosmic gas. By a redshift of $z \sim 5$, the average metal abundance in the IGM is observed to be $\sim 1\%$ of the solar value, as expected from the heavy element yield of the same massive stars that reionized the Universe.[31]

The initial mass function of metal-free stars was likely affected by radiative feedback before it was influenced by metal enrichment feedback. Early on, ionizing photons from the first stars are expected to travel farther than the metal-rich outflows from their associated supernovae. It is therefore likely that a second generation of stars with an intermediate characteristic mass ($\sim 10M_{\odot}$) formed in metal-free regions that were photo-ionized by the very first stars, inside of which molecular chemistry was accelerated and cooling by H_2 and HD was enhanced.[32]

5.2.3 Properties of the First Stars

Primordial stars that are hundreds of times more massive than the Sun 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, $\sim 5800\,{\rm K}$. These massive stars are held against their self-gravity by radiation pressure, having the so-called *Eddington luminosity* (see derivation in §6.3) which is strictly proportional to their mass $m_\star$,

$$L_E = 1.3 \times 10^{40} \left(\frac{m_\star}{100 M_\odot}\right)\ {\rm erg\ s^{-1}}, \tag{5.5}$$

and is a few million times 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 calculated 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

$$R_\star = \left(\frac{L_E}{4\pi\sigma T_{\rm eff}^4}\right)^{1/2} = 4.3 \times 10^{11}\ {\rm cm} \times \left(\frac{m_\star}{100 M_\odot}\right)^{1/2}, \tag{5.6}$$

which is ~ 6 times larger than the radius of the Sun, $R_\odot = 7 \times 10^{10}\ {\rm cm}$.

The high surface temperature of the first stars makes them ideal factories of ionizing photons. To free (ionize) an electron out of a hydrogen atom requires an energy of 13.6eV (equivalent, through a division by Boltzmann's constant k_B, to a temperature of $\sim 1.6 \times 10^5$K), which is coincidentally the characteristic energy of a photon emitted by the first massive stars.

The first stars had lifetimes of a few million years, independent of their mass. During its lifetime, a Population III star produced $\sim\ 10^5$ ionizing photons per proton incorporated in it. 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. The actual required star formation efficiency depends on the fraction of all ionizing photons that escape from the host galaxies into the intergalactic space ($f_{\rm esc}$) rather then being absorbed by hydrogen inside these galaxies.[33] For comparison, Population II stars produce on average $\sim\ 4,000$ ionizing photons per proton in them. If the Universe was ionized by such stars, then a much larger fraction (by a factor of $\sim\ 25$) of its gaseous content had to be converted into stars in order to have the same effect as Population III stars.

The first stars had a large impact on their gaseous environment. Their UV emission ionized and heated the surrounding gas, and their winds or supernova explosions pushed the gas around like a piston. Such feedback effects controlled the overall star formation efficiency within each galaxy, $f_\star$. This efficiency is likely to have been small in the earlier galaxies that had shallow potential wells, and it is possible that the subsequent Population II stars dominated the production of ionizing photons during reionization. By today, the global fraction of baryons converted into stars in the Universe is $\sim 10\%$.[34]

The accounting of the photon budget required for reionization is simple. If only a fraction $f_\star \sim 10\%$ of the gas in galaxies was converted into Population II stars and only $f_{\rm esc} \sim 10\%$ of the ionizing radiation escaped into intergalactic space, then more than $\sim 1/(4000 \times 10\% \times 10\%) = 2.5\%$ of the matter in the Universe had to collapse into galaxies before there was one ionizing photon available per intergalactic hydrogen atom. Reionization completed once the number of ionizing photons grew by another factor of a few to compensate for recombinations in dense intergalactic regions.

It is also possible to predict the luminosity distribution of the first galaxies as a function of redshift and photon wavelength by "dressing up" the mass distribution of halos in Figure 3.2 with light. The simplest prescription would be to assume that some fraction $f_\star(\Omega_b/\Omega_m)$ of the total mass in each halo above the Jeans mass is converted into stars with a prescribed stellar mass distribution (Population II or Population III) over some prescribed period of time (related to the rotation time of the disk). Using available computer codes for the combined spectrum of the stars as a function of time, one may then compute the luminosity distribution of the halos as a function of redshift and wavelength and make predictions for future observations.[35] The association of specific halo masses with galaxies of different luminosities can also be guided by their clustering properties.[36]

The end state in the evolution of massive Population III stars depends on their mass and rotation rate. Ignoring rotation, one finds that within an intermediate mass range of 140–$260 M_\odot$ they were likely to explode as energetic *pair-instability supernovae*, and outside this mass range they were likely to implode into black holes.[37] A pair-instability supernova is triggered when the core of the very massive low-metallicity star heats up in the last stage of its evolution. This 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. 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 (§12.1.4).

Where are the remnants of the first stars located today? The very first stars formed in rare high-density peaks, hence their black hole remnants are likely to populate the cores of present-day galaxies. However, the bulk of the stars which formed in low-mass systems at later times are expected to behave similarly to the collisionless dark matter particles, and populate galaxy halos.

Although the very first generation of local galaxies are buried deep in the core of the Milky Way, most of the stars there today formed much later, making the search for rare old stars as impractical as finding needles in a haystack. Because the outer Milky Way halo is far less crowded with younger stars, it is much easier to search for old stars there. Existing halo surveys discovered a population of stars with exceedingly low iron abundance ($\sim 10^{-5}$ of the solar abundance of iron relative to hydrogen), but these "anemic" stars have a high abundance of other heavy elements, such as carbon.[38] We do not expect to find the very first population of massive stars in these surveys, since these had a lifetime of only a few million years, several orders of magnitude shorter than the period of time that has elapsed since the dark ages.

5.2.4 Feedback (UV Illumination, Metal Enrichment, Remnants)

5.3 LATER GENERATIONS OF STARS

5.4 GLOBAL PARAMETERS OF HIGH-REDSHIFT GALAXIES

5.4.1 Minimum Mass

5.4.2 Size Distribution

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,

$$\lambda \equiv J|E|^{1/2}G^{-1}M^{-5/2} \; . \tag{5.7}$$

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 beyond which the halo would break up. As the baryons cool and lose their pressure support, they settle to a rotationally-supported disk, whose mass is a fraction m_d of the halo mass and its angular momentum is a fraction j_d of that of the halo. The scale radius of the disk is set by rotational support,[39]

$$R_d = \frac{1}{\sqrt{2}}\left(\frac{j_d}{m_d}\right)\lambda\, r_{\rm vir} \; . \tag{5.8}$$

At a fixed halo mass, the size of the associated disk is expected to decrease with increasing redshift at $z \gg 1$ as $R_d \propto (1+z)^{-1}$. Observations indicate that the luminous cores of galaxies follow this expected trend over the wide redshift range of $2 < z < 8$, as illustrated in Figure 5.6.

The observed distribution of disk sizes in local galaxies suggests that the specific angular momentum of the disk is similar to that of the halo, and so we assume $j_d/m_d = 1$. The distribution of disk sizes is then determined by the halo abundance and by the distribution of spin parameters. N-body simulations [40] indicate that the latter approximately follows a lognormal distribution,

$$p(\lambda)d\lambda = \frac{1}{\sigma_\lambda\sqrt{2\pi}}\exp\left[-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2}\right]\frac{d\lambda}{\lambda} \; , \tag{5.9}$$

with $\bar{\lambda} = 0.05$ and $\sigma_\lambda = 0.5$.

The distribution of disks is truncated at the low-mass end due to the fact that gas pressure inhibits baryon collapse and disk formation in shallow potential wells, i.e.

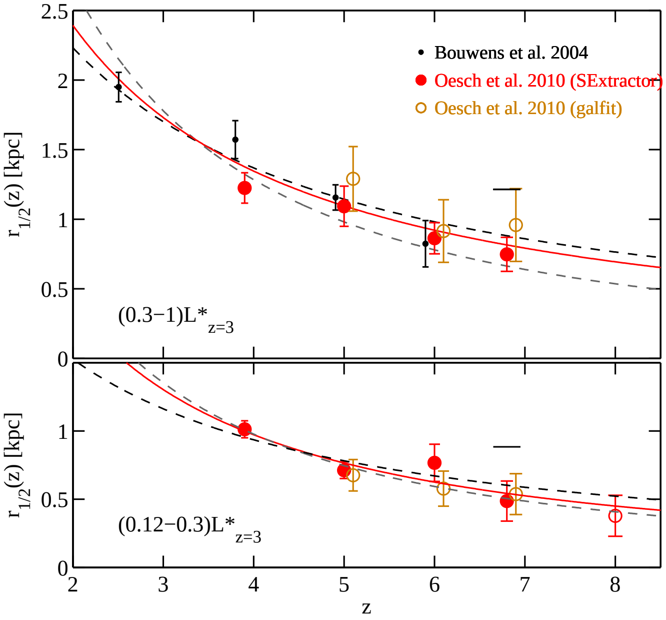

Figure 5.6 Observed evolution of the mean half-light radius of galaxies across the redshift range $2 < z < 8$ in two bins of fixed intrinsic luminosity: (0.3-1)$L_*(z = 3)$ (top) and (0.12-0.3)$L_*(z = 3)$ (bottom), where $L_*(z = 3)$ is the characteristic luminosity of a galaxy at $z = 3$ (Eq. 12.6). 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).

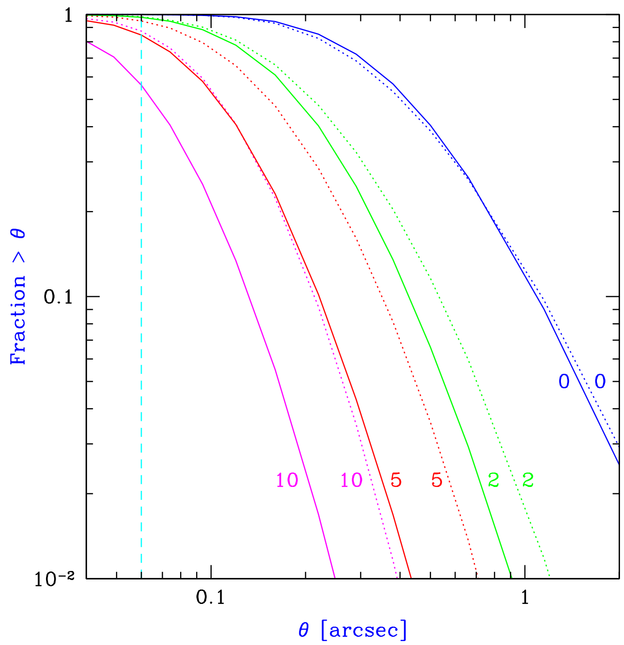

Figure 5.7 Theoretically predicted distribution of galactic disk sizes in various redshift intervals. Each curve shows the fraction of the total number counts contributed by sources larger than an observed angle θ in arcseconds. The diameter θ is measured out to one scale length, R_d. The label next to each curve indicates the lower limit of the redshift interval; i.e. '0' indicates sources with $0 < z < 2$, and so forth for sources with $2 < z < 5$, $5 < z < 10$, and $z > 10$. Figure credit: Barkana, R., & Loeb, A. *Astrophys. J.* **531**, 613 (2000).

in halos with a low circular velocity V_c. In particular, photo-ionization heating by the cosmic UV background heats the intergalactic gas to a characteristic temperature of $\sim 10^4$ K and prevents it from settling into systems with a virial temperature below $\sim 10^5$K.

Figure 5.7 shows the fraction of the total number counts contributed by sources with diameters greater than θ, as a function of θ, for a minimum circular velocity of $V_{\rm cut} = 50\ {\rm km\ s^{-1}}$.

Observationally, the star formation rate per unit area in galaxies is correlated (over $\sim$ 7 orders of magnitude) with the total surface mass density of molecular and atomic gas to the power of $\sim 1.4 \pm 0.1$. This so-called *Kennicutt-Schmidt relation*[41] can also be interpreted in terms of a fixed fraction of the cold gas being converted into stars per dynamical time in the associated galactic disks. 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.

5.5 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 view 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.[42]

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-

[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).

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 form and explode only a million years after being born. 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[43] (see illustration 5.8). 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 6.3. However, when jets are generated in the core of a star, they have to 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 happened to line up with its 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 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[44] at a redshift 8.2, representing the most distant source known, originating at the time when the Universe was only ~ 630 million years old.[45]

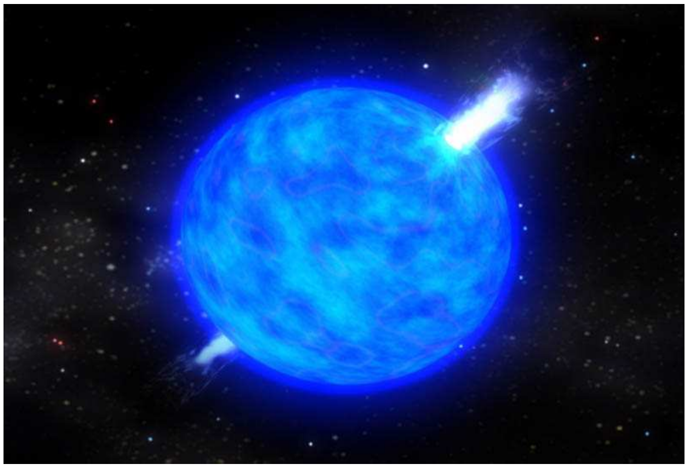

Figure 5.8 Illustration of a long-duration gamma-ray burst in the popular “collapsar” model. The collapse of the core of a massive star (which lost its hydrogen envelope) to a black hole generates two opposite jets moving out at a speed close to the speed of light. The jets drill a hole in the star and 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

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, tidal torques 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.

We start this chapter with a short introduction to the properties of black holes in general relativity.

6.1 BASIC PRINCIPLES OF ASTROPHYSICAL BLACK HOLES

Birkhoff's theorem states that the only vacuum, spherically symmetric gravitational field is that described by the static *Schwarzschild metric*,

$$ds^2 = -\left(1 - \frac{r_{\rm Sch}}{r}\right) c^2 dt^2 + \left(1 - \frac{r_{\rm Sch}}{r}\right)^{-1} dr^2 + r^2 d\Omega, \tag{6.1}$$

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,

$$r_{\rm Sch} = \frac{2GM}{c^2} = 2.95 \times 10^5 \text{ cm} \left(\frac{M}{1 M_\odot}\right). \tag{6.2}$$

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,

$$ds^2 = -\left(1 - \frac{r_{\rm Sch} r}{\Sigma_k}\right) c^2 dt^2 - \frac{2 j r_{\rm Sch} r \sin^2\theta}{\Sigma_k} c dt d\phi + \frac{\Sigma_k}{\Delta} dr^2$$
$$+ \Sigma_k d\theta^2 + \left(r^2 + j^2 + \frac{r_{\rm Sch} j^2 r \sin^2\theta}{\Sigma_k}\right)\sin^2\theta d\phi^2. \qquad (6.3)$$

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.

Test particles 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,

$$r_{\rm ph} = r_{\rm Sch}\left[1 + \cos\left(\frac{2}{3}\cos^{-1}[\pm a]\right)\right], \qquad (6.4)$$

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 $2r_{\rm Sch}$ 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

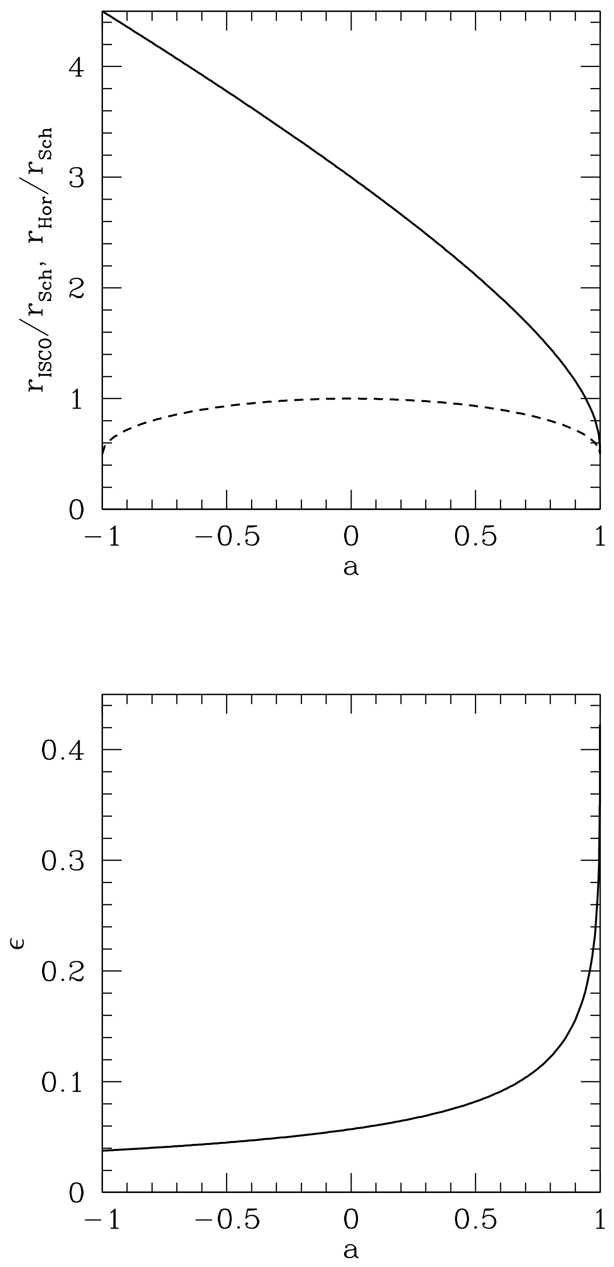

Figure 6.1 The left panel shows the radius of the black hole horizon $r_{\rm Hor}$ (dashed line) and the *Innermost Circular Stable Orbit (ISCO)* around it $r_{\rm ISCO}$ (solid line), in units of the Schwarzschild radius $r_{\rm Sch}$ (see equation 6.2), as functions of the black hole spin parameter 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.

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[46],

$$r_{\rm ISCO} = \frac{1}{2} r_{\rm Sch} \left\{ 3 + Z_2 \pm [(3-Z_1)(3+Z_1+2Z_2)]^{1/2} \right\}, \qquad (6.5)$$

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 6.1 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,

$$\epsilon = 1 - \frac{r^2 - r_{\rm Sch} r \mp j\sqrt{\frac{1}{2} r_{\rm Sch} r}}{r(r^2 - \frac{3}{2} r_{\rm Sch} r \mp 2j\sqrt{\frac{1}{2} r_{\rm Sch} r})^{1/2}}. \qquad (6.6)$$

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.

6.2 ACCRETION OF GAS ONTO BLACK HOLES

6.2.1 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,

$$\dot{M} \approx \pi r_{\rm inf}^2 \rho_0 c_s = 15 \left(\frac{M}{10^8 M_\odot}\right)^2 \left(\frac{n_0}{1~{\rm cm}^{-3}}\right) \left(\frac{T_0}{10^4~{\rm K}}\right)^{-3/2} M_\odot~{\rm yr}^{-1}. \qquad (6.7)$$

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 (6.7). 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 (6.7).

6.2.2 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 $\sim 10^{12}\,\mathrm{K}(r/r_{\mathrm{Sch}})^{-1}$. 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),[47] 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 $v_r \ll v_\phi$ 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,

$$\dot{M} = 2\pi r \Sigma v_r = const, \tag{6.8}$$

where $\Sigma(r)$ is the surface mass density of the disk and v_r is the radial (accretion) velocity of the gas.

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.

Within a Keplerian accretion disk, the flux 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 6.1. Angular momentum conservation requires that the net rate of its change within a radius r be equal to the viscous torque, namely

$$f_\phi \times (2\pi r \times 2h) \times r = \dot{M}\left[(GMr)^{1/2} - (GMr_{\rm ISCO})^{1/2}\right]. \tag{6.9}$$

The production rate of heat by the viscous stress is given by $\dot{Q} = f_\phi^2/\eta$. Substituting f_ϕ and equation (6.9) gives,

$$2h\dot{Q} = \frac{3\dot{M}}{4\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{6.10}$$

This power gives local flux that is radiated vertically from the top and bottom surfaces of the disk,

$$F = \frac{1}{2} \times 2h\dot{Q} = \frac{3\dot{M}}{8\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{6.11}$$

The total luminosity of the disk is given by,

$$L = \int_{r_{\rm ISCO}}^{\infty} 2F \times 2\pi r dr = \frac{1}{2}\frac{GM\dot{M}}{r_{\rm ISCO}}, \tag{6.12}$$

where we have ignored genral-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,

$$\frac{1}{\rho}\frac{dP}{dz} = -\frac{GM}{r^2}\frac{z}{r}, \tag{6.13}$$

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. Instead the magneto-rotational instability[48] is likely to develop turbulent eddies in the disk which are much more effective at transporting its angular momentum. 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,

$$\beta \equiv \frac{P_{\rm gas}}{P}, \tag{6.14}$$

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,

$$F = -c \frac{dP_{\rm rad}}{d\tau}, \tag{6.15}$$

where the optical-depth τ is related to the frequency-averaged (so-called, Rosseland-mean) opacity coefficient of the gas, κ,

$$\tau = \int_0^h \kappa \rho dz \approx \frac{1}{2} \kappa \Sigma, \tag{6.16}$$

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

$$\kappa_{\rm es} = \frac{\sigma_{\rm T}}{m_p} = 0.4\,{\rm cm}^2\,{\rm g}^{-1}, \tag{6.17}$$

and *free-free* absorption with

$$\kappa_{\rm ff} \approx 8 \times 10^{22} {\rm cm}^2\,{\rm g}^{-1} \left(\frac{\rho}{{\rm g\,cm}^{-3}} \right) \left(\frac{T}{\rm K} \right)^{-7/2}, \tag{6.18}$$

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 6.33 below). When the luminosity approaches the Eddington limit, the disk bloates and h approaches r, violating the thin-disk assumption. We write $\dot{m} = (\dot{M}/\dot{M}_E)$, 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 6.11) 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,

$$T_d \approx \left(\frac{4F}{a} \right)^{1/4} = 10^5\ {\rm K}\ M_8^{-1/4} \dot{m}_{-1}^{1/4} r_1^{-3/4} \left[1 - \left(\frac{r}{r_{\rm ISCO}} \right)^{1/2} \right]. \tag{6.19}$$

The accretion disk can be divided radially into three distinct regions, [49]

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

$$r_1 \approx 54\, \alpha_{-1}^{2/21} (\dot{m}_{-1}/\epsilon_{-1})^{16/21} M_8^{2/21} \quad \text{if b} = 1, \tag{6.20}$$

$$58\, \alpha_{-1}^{2/21} (\dot{m}_{-1}/\epsilon_{-1})^{16/21} M_8^{2/21} \quad \text{if b} = 0, \tag{6.21}$$

and the transition radius between regions 2 and 3 is

$$r_1 \approx 4 \times 10^2\, (\dot{m}_{-1}/\epsilon_{-1})^{2/3}. \tag{6.22}$$

The surface density and scale-height of the disk are given by,
Inner region:

$$\Sigma(r) \approx (3 \times 10^6\, \mathrm{g\, cm^{-2}}) \alpha_{-1}^{-4/5} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5} \ \text{ if b} = 1, \tag{6.23}$$

$$(8 \times 10^2\, \mathrm{g\, cm^{-2}}) \alpha_{-1}^{-1} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{-1} r_1^{3/2} \quad \text{if b} = 0, \tag{6.24}$$

$$h(r) \approx R_{\mathrm{Sch}} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right). \tag{6.25}$$

Middle region:

$$\Sigma(r) \approx (3 \times 10^6\, \mathrm{g\, cm^{-2}}) \alpha_{-1}^{-4/5} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5}, \tag{6.26}$$

$$h(r) \approx 1.4 \times 10^{-2} R_S \alpha_{-1}^{-1/10} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{1/5} M_8^{-1/10} r_1^{21/20}. \tag{6.27}$$

Outer region:

$$\Sigma(r) \approx (6 \times 10^6\, \mathrm{g\, cm^{-2}}) \alpha_{-1}^{-4/5} \left(\frac{\dot{m}_{-1}}{\epsilon_{0.1}}\right)^{7/10} M_8^{1/5} r_1^{-3/4}, \tag{6.28}$$

$$h(r) \approx 10^{-2} R_S \alpha_{-1}^{-1/10} \left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/20} M_8^{-1/10} r_1^{9/8}. \tag{6.29}$$

The mid-plane temperature is given by,

$$T_m(r) \approx \left(16\pi^2\right)^{-1/5} \left(\frac{m_p}{k_B \sigma_T}\right)^{1/5} \alpha^{-1/5} \kappa^{1/5} \dot{M}^{2/5} \Omega^{3/5} \beta^{-(1/5)(b-1)}. \tag{6.30}$$

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.[50] For the above scaling laws of the outer disk, this occurs at the outer radius,

$$r_1 \approx 2 \times 10^4 \alpha_{-1}^{28/45} \left(\dot{m}_{-1}/\epsilon_{-1}\right)^{-22/45} M_8^{52/45}. \tag{6.31}$$

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.

6.2.3 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[51] 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 anayltic RIAF models typically admit a radiative efficiency of,

$$\frac{L}{\dot{M}c^2} \approx \epsilon \left(\frac{\dot{M}}{\dot{M}_{\rm crit}} \right), \tag{6.32}$$

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

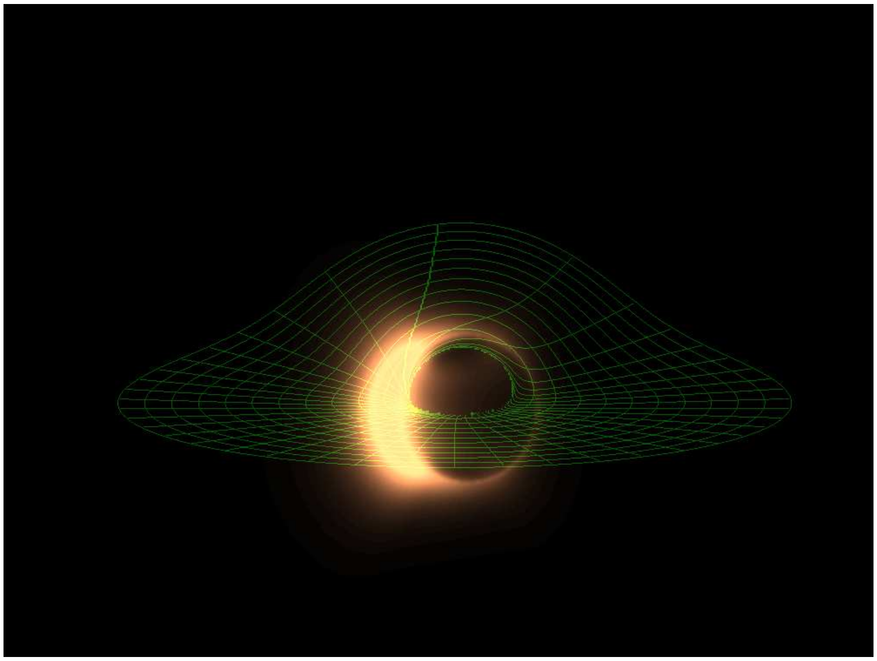

Figure 6.2 Simulated image of an accretion flow around a black hole spinning at half its maximum rate, from a viewing angle of 10° relative to the rotation axis. The coordinate grid in the equatorial plane of the spiraling flow shows how strong lensing around the black hole bends the back of the apparent 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).

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.

6.3 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 6.3). 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

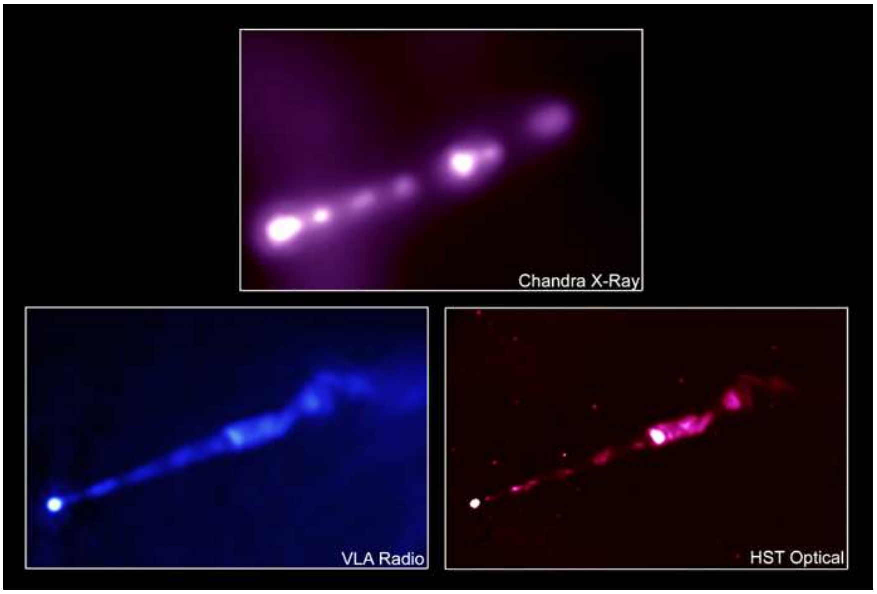

Figure 6.3 Multi-wavelength images of the highly collimated jet emanating from the supermassive black hole at the center of the giant elliptical galaxy M87. The X-ray image (top) was obtained with the Chandra X-ray satellite, the radio image (bottom left) was obtained with the Very Large Array (VLA), and the optical image (bottom right) was obtained with the Hubble Space Telescope (HST).

distance between galaxies.[52] 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 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 §6.1 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 6.1). 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.[53] 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 Eddington luminosity, L_E, is obtained from 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:

$$\frac{GM_{\rm BH}m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{6.33}$$

where m_p is the proton mass and $\sigma_T = 0.67 \times 10^{-24}\ {\rm cm}^2$ 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 to the luminosity of massive stars which are held at rest against gravity by radiation pressure, as described by equation (6.34). This limit is formally valid in a spherical

[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 (corresponding 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).

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,

$$L_E = 1.3 \times 10^{44} \left(\frac{M_{\rm BH}}{10^6 M_\odot} \right) \text{ erg s}^{-1}, \tag{6.34}$$

Interestingly,[54] 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,[55] 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,

$$\int \sigma(\nu) d\nu = f_{12} \left(\frac{\pi e^2}{m_e c} \right), \tag{6.35}$$

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.

The total luminosity from gas accreting onto a black hole, L, can be written as some radiative efficiency ϵ times the mass accretion rate $\dot{M}$,

$$L = \epsilon \dot{M} c^2, \tag{6.36}$$

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

$$\dot{M}_{\rm BH} = \frac{M_{\rm BH}}{t_E}, \tag{6.37}$$

where (after substituting all fundamental constants),

$$t_E = 4 \times 10^7 \text{years} \left(\frac{\epsilon/(1-\epsilon)}{10\%} \right) \left(\frac{L}{L_E} \right)^{-1}. \tag{6.38}$$

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 (6.38) 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.[56] 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).[57] 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.[58] This implies that continuous accretion was likely punctuated by black hole ejection events,[59] 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 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 §3.1. 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 6.4 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.

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

[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)).

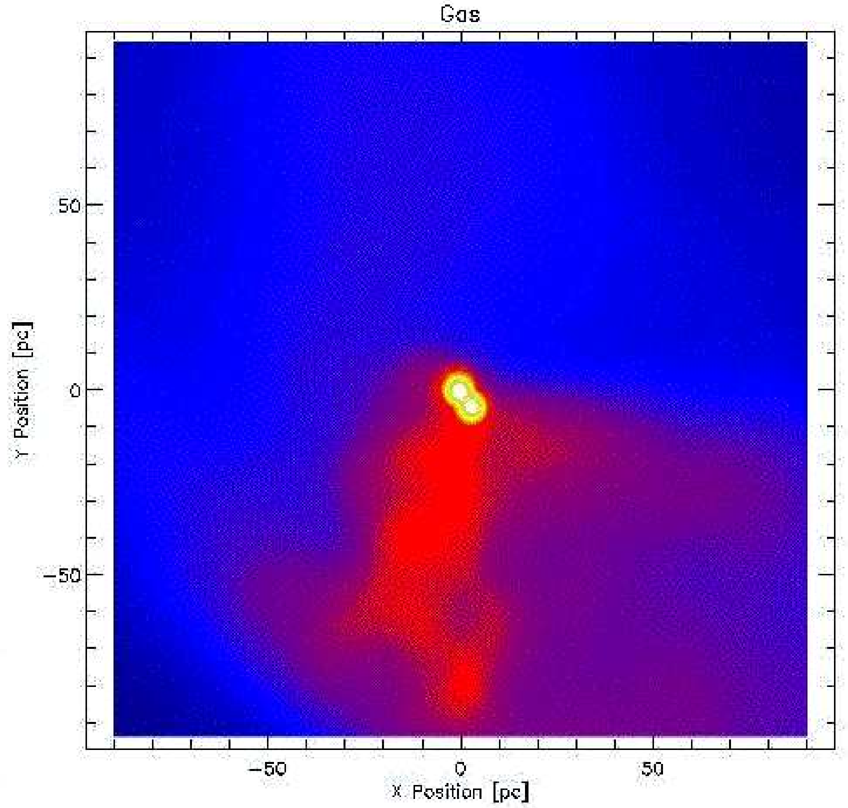

Figure 6.4 Numerical simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2. The image shows a snapshot of the gas density distribution 500 million years after the Big Bang, indicating the formation of two compact objects 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).

tight correlations between the mass of central black holes in present-day galaxies and the velocity dispersion[60] $\sigma_\star$ or luminosity[61] $L_{\rm sp}$ 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 (3.29), 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 properties of their host galaxies.[62] 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[63]. This explains the high abundance of heavy elements inferred from the broad emission lines of quasars at all redshifts[64].

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 modelling galaxies, it is possible to “dress up” the mass distribution of halos in Figure 3.2 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), and find the evolution of the quasar population over redshift. This simple approach can be tuned to give good agreement with existing data on quasar evolution.[65]

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.[66] 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 6.2).

6.4 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[67] $\sim \varepsilon^{0.4}$. 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,

$$a_{\rm hard} \approx 0.1 \frac{q}{(1+q)^2} M_6 \left(\frac{\sigma_\star}{100\ {\rm km\ s^{-1}}} \right)^{-2}\ {\rm pc}, \tag{6.39}$$

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,

$$P = 2\pi (GM/a^3)^{-1/2} = 1.72 \times 10^{-2}\ {\rm yr}\ a_{14}^{3/2} M_6^{-1/2}, \tag{6.40}$$

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

$$v = v_1 + v_2 = (2\pi a/P) = 1.15 \times 10^4\ {\rm km\ s^{-1}} M_6^{1/2} a_{14}^{-1/2}\ . \tag{6.41}$$

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%, $\dot{M}_E = 0.023 M_\odot\ \mathrm{yr}^{-1} M_6$. We then find,

$$t_{\rm gas} \approx (J/\dot{M}va) = \mu/\dot{M} = 1.1 \times 10^7\ \mathrm{yr}\ \zeta \dot{\mathcal{M}}^{-1}. \tag{6.42}$$

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,

$$t_{\rm GW} = \frac{5}{256} \frac{c^5 a^4}{G^3 M^2 \mu} = 2.53 \times 10^3\ \mathrm{yr}\ \frac{a_{14}^4}{\zeta M_6^3}. \tag{6.43}$$

By setting $t_{\rm GW} = t_{\rm gas}$ we can solve for the orbital speed, period, and separation at which GWs take over,

$$v_{\rm GW} = 4.05 \times 10^3\ \mathrm{km\ s}^{-1}\ \zeta^{-1/4} (\dot{\mathcal{M}} M_6)^{1/8}\ ; \tag{6.44}$$

$$P_{\rm GW} = 0.4\ \mathrm{yr}\ \zeta^{3/4} M_6^{5/8} \dot{\mathcal{M}}^{-3/8}\ ; \tag{6.45}$$

$$a_{\rm GW} = 2.6 \times 10^{-4}\ \mathrm{pc}\ \zeta^{1/2} M_6^{3/4} \dot{\mathcal{M}}^{-1/4}. \tag{6.46}$$

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 $\sim 200~\mathrm{km~s^{-1}}$ in a merger of non-spinning black holes, and up to $\sim 4,000~\mathrm{km~s^{-1}}$ 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.[68] 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.

Chapter Seven

The Reionization of Cosmic Hydrogen by the First Galaxies

7.1 IONIZATION SCARS BY THE FIRST STARS

The cosmic microwave background (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. Observations 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. Indeed, the bulk of the Universe's ordinary matter today is in the form of free electrons and protons, located deep in intergalactic space. The free electrons have other side effects; for example, they scatter the CMB and produce polarization fluctuations at large angles on the sky.[i] The latest analysis of the CMB polarization data from WMAP[69] indicates that $8.7 \pm 1.7\%$ of the CMB photons were scattered by free electrons after cosmological recombination, implying that reionization took place at around a redshift $z \sim 10$, only 500 million years after the Big Bang.[ii] 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 can address this question using the formation rate of new galaxy halos at various cosmic epochs. 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 above redshift 6, indicating that stars most likely dominated the production of ionizing UV photons during the reionization

[i] Polarization is produced when free electrons scatter a radiation field with a quadrupole anisotropy Q. Consequently, reionization generates a fractional CMB polarization of $P \sim 0.1\tau Q$ out of the CMB quadrupole on scales of the horizon at reionization. Here, $\tau = \int_0^{z_{\rm reion}} n_e(z)\sigma_T(cdt/dz)dz$ is the optical depth for scattering by electrons of mean density $n_e(z)$, which provides a measure of the redshift of reionization, $z_{\rm reion}$. (The scattering probability, $\tau \ll 1$, is the chance that a photon will encounter an electron within the volume associated with the scattering cross-section σ_T times the photon path length $\int cdt$.) For a sudden reionization of hydrogen and neutral helium at redshift $z_{\rm reion}$, $\tau = 4.75 \times 10^{-3} \times \{[\Omega_\Lambda + \Omega_m(1+z_{\rm reion})^3]^{1/2} - 1\}$.

[ii] This number will be refined by forthcoming CMB data from the Planck satellite (http://www.rssd.esa.int/index.php?project=planck), and will be supplemented by constraints from 21-cm observations [see Pritchard, J., Loeb, A., & Wyithe, J. S. B., http://arxiv.org/abs/0908.3891 (2009)].

epoch.[70] Since 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 inter-galactic medium (IGM) during reionization a two-phase medium characterized by highly ionized regions separated from neutral regions by sharp ionization fronts. We can obtain a first estimate of the requirements of reionization by demanding one stellar ionizing photon for each hydrogen atom in the Universe. If we conservatively assume that stars within the early galaxies were similar to those observed locally, then each star produced $\sim 4,000$ ionizing photons per proton in it. Star formation is observed today to be an inefficient process, but even if stars in galaxies formed out of only a fraction $f_\star \sim 10\%$ of the available gas, this amount was still sufficient to assemble only a small fraction of the total mass in the universe into galaxies in order to ionize the entire IGM. This fraction would be $\sim 2.5\%(f_{\rm esc}/10\%)^{-1}$ for an escape fraction $f_{\rm esc}$ of ionizing UV photons out of galaxies. More detailed estimates of the actual required fraction account for the formation of some Population III stars (which were more efficient ionizers), and for recombinations of hydrogen atoms at high redshifts and in dense regions. At the mean IGM density, the recombination time of hydrogen is shorter than the age of the Universe at $z > 8$.

From studies of the processed spectra of quasars at $z \sim 6$, we know that the IGM is highly ionized a billion years after the Big Bang. There are hints, however, 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. We now know that the universe could not have fully reionized earlier than an age of 300 million years, since WMAP observed the effect of the freshly created plasma at reionization on the large-scale polarization anisotropies of the CMB which limits the reionization redshift; an earlier reionization, when the universe was denser, would have created a stronger scattering signature that would have been inconsistent with the WMAP observations. In any case, the redshift at which reionization ended only constrains the overall cosmic efficiency for producing ionizing photons. In comparison, a detailed picture of reionization in progress will teach us a great deal about the population of the first galaxies that produced this cosmic phase transition.

7.2 PROPAGATION OF IONIZATION FRONTS

Astronomers label the neutral and singly ionized states of an atomic species as I and II; for example, abbreviating neutral hydrogen as H I and ionized hydrogen as H II. The radiation output from the first stars ionizes H I in a growing volume, eventually encompassing almost the entire IGM within a single H II bubble. In the early stages of this process, each galaxy produced a distinct H II region, and only when the overall H II filling factor became significant did neighboring bubbles begin to overlap in large numbers, ushering in the “overlap phase” of reionization. Thus, the first goal of a model of reionization is to describe the initial stage, during which each source produces 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, each hydrogen atom in the IGM would only have to be ionized once, and the ionized physical volume V_p would simply be determined by

$$\bar{n}_H V_p = N_\gamma \ , \tag{7.1}$$

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. However, the elevated density of the IGM at high redshift implies that recombinations cannot be ignored. Just before World War II, the Danish astronomer Bengt Strömgren analyzed the same problem for hot stars embedded in the interstellar medium.[71] 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:

$$\alpha_B \bar{n}_H^2 V_p = \frac{d\,N_\gamma}{dt} \ , \tag{7.2}$$

where the recombination rate depends on the square of the density and on the recombination coefficient[iii] (to all states except the ground energy level, which would just recycle the ionizing photon) $\alpha_B = 2.6 \times 10^{-13}$ cm^3 s^{-1} for hydrogen at $T = 10^4$ K. The complete description of the evolution of an expanding H II region, including the ingredients of a non-steady ionizing source, recombinations, and cosmological expansion, is given by[72]

$$\bar{n}_H \left(\frac{dV_p}{dt} - 3HV_p \right) = \frac{d\,N_\gamma}{dt} - \alpha_B \left\langle n_H^2 \right\rangle V_p \ . \tag{7.3}$$

In this equation, the mean density $\bar{n}_H$ varies with time as $1/a^3(t)$. Note that the recombination rate scales as the square of the density. Therefore, if the IGM is not uniform, but instead the gas which is being ionized is mostly distributed in high-density clumps, then the recombination time will be shorter. This is often accommodated for by introducing a volume-averaged clumping factor C (which is, in general, time dependent), defined by[iv]

$$C = \left\langle n_H^2 \right\rangle / \bar{n}_H^2 \ . \tag{7.4}$$

If the ionized volume is large compared to the typical scale of clumping, so that many clumps are averaged over, then equation (7.3) can be solved by supplementing it with equation (7.4) and specifying C. Switching to the comoving volume V, the resulting equation is

$$\frac{dV}{dt} = \frac{1}{\bar{n}_H^0} \frac{d\,N_\gamma}{dt} - \alpha_B \frac{C}{a^3} \bar{n}_H^0 V \ , \tag{7.5}$$

where the present number density of hydrogen is

$$\bar{n}_H^0 = 2.1 \times 10^{-7} \ \mathrm{cm}^{-3} \ . \tag{7.6}$$

[iii] See §2 in Osterbrock, D. E., & Ferland, G. J. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*, University Science Books, Sausalito (2006).

[iv] The recombination rate depends on the number density of electrons, and in using equation (7.4) we are neglecting the small contribution made by partially or fully ionized helium.

This number density is lower than the total number density of baryons $\bar{n}_b^0$ by a factor of ~ 0.76, corresponding to the primordial mass fraction of hydrogen. The solution for $V(t)$ around a source which turns on at $t = t_i$ is [73]

$$V(t) = \int_{t_i}^{t} \frac{1}{\bar{n}_H^0} \frac{dN_\gamma}{dt'} e^{F(t',t)} dt' \ , \tag{7.7}$$

where

$$F(t',t) = -\alpha_B \bar{n}_H^0 \int_{t'}^{t} \frac{C(t'')}{a^3(t'')} dt'' \ . \tag{7.8}$$

At high redshifts ($z \gg 1$), the scale factor varies as

$$a(t) \simeq \left(\frac{3}{2} \sqrt{\Omega_m} H_0 t \right)^{2/3} \ , \tag{7.9}$$

and with the additional assumption of a constant C, the function F simplifies as follows. Defining

$$f(t) = a(t)^{-3/2} \ , \tag{7.10}$$

we derive

$$F(t',t) = -\frac{2}{3} \frac{\alpha_B \bar{n}_H^0}{\sqrt{\Omega_m} H_0} C \left[f(t') - f(t) \right] = -0.26 \left(\frac{C}{10} \right) \left[f(t') - f(t) \right] \ . \tag{7.11}$$

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 of $f_\star = 10\%$, and that the escape fraction for the resulting ionizing radiation is also $f_{\rm esc} = 10\%$. As mentioned in §5.2.2, Population II stars produce a total of $N_\gamma \approx 4,000$ ionizing photons per baryon in them. We may define a parameter which gives the overall number of ionizations per baryon assembled into the halo,

$$N_{\rm ion} \equiv N_\gamma \, f_\star \, f_{\rm esc} \ . \tag{7.12}$$

If we neglect recombinations then we obtain the maximum comoving radius of the region which the halo of mass M can ionize as,

$$\begin{aligned} r_{\rm max} = \left(\frac{3}{4\pi} \frac{N_\gamma}{\bar{n}_H^0} \right)^{1/3} &= \left(\frac{3}{4\pi} \frac{N_{\rm ion}}{\bar{n}_H^0} \frac{\Omega_b}{\Omega_m} \frac{M}{m_p} \right)^{1/3} \\ &= 680 \, {\rm kpc} \left(\frac{N_{\rm ion}}{40} \frac{M}{10^8 M_\odot} \right)^{1/3} \ . \end{aligned} \tag{7.13}$$

However, the actual radius never reaches this size if the recombination time is shorter than the lifetime of the ionizing source.

We may obtain a similar estimate for the size of the H II region around a galaxy if we consider a quasar rather than stars. 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 the collapsed fraction of 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. If we take the limit of an extremely bright source, characterized by an arbitrarily high production rate of ionizing photons, then equation 7.3 would imply that the H II region expands faster than light. This result is clearly unphysical and must be corrected for bright sources. At early times, the ionization front around a bright quasar would have expanded at nearly the speed of light, c, but this only occurs when the H II region is sufficiently small such that the production rate of ionizing photons by the central source exceeds their consumption rate by hydrogen atoms within this volume. It is straightforward to do the accounting for these rates (including recombination) by taking the light propagation delay into account. The general equation for the relativistic expansion of the *comoving* radius $R = (1+z)r_{\rm p}$ of a quasar H II region in an IGM with neutral filling fraction $x_{\rm HI}$ (fixed by other ionizing sources) is given by[74],

$$\frac{dR}{dt} = c(1+z)\left[\frac{\dot{N}_\gamma - \alpha_{\rm B} C x_{\rm HI}\left(\bar{n}_{\rm H}^0\right)^2 (1+z)^3 \left(\frac{4\pi}{3}R^3\right)}{\dot{N}_\gamma + 4\pi R^2 (1+z)\, c x_{\rm HI} \bar{n}_{\rm H}^0}\right], \tag{7.14}$$

where $\dot{N}_\gamma$ is the rate of ionizing photons crossing a shell of the H II region at radius R and time t. Indeed, for $\dot{N}_\gamma \to \infty$ the propagation speed of the physical radius of the H II region $r_p = R/(1+z)$ approaches the speed of light in the above expression, $(dr_p/dt) \to c$.

The process of the reionization of hydrogen involves several distinct stages.[75] 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.

The central, relatively rapid "overlap" phase of reionization begins 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. Since each bubble coalescence accelerates the process of reionization, the overlap phase has the character of a phase transition and is expected to occur rapidly. By the end of this stage, most regions in the IGM are able to "see" several unobscured sources, therefore the ionizing intensity is much higher and more homogeneous than before overlap. An additional attribute of this rapid "overlap" phase results from the fact that hierarchical structure formation models predict a galaxy formation rate that rises rapidly with time at these high redshifts. This is because most galaxies fall on the exponential tail of the Press-Schechter mass function at early cosmic times, and so the fraction of mass that gets incorporated into stars grows exponentially with time. This process leads to a final state in which the low-density IGM has been highly ionized, with ionizing radiation reaching everywhere except gas located inside self-shielded, high-density clouds. This "moment of reionization" marks the end of the "overlap phase."

Some neutral gas does, however, remain in high-density structures, which are gradually ionized as galaxy formation proceeds and the mean ionizing intensity grows with time. The ionizing intensity continues to grow and become more uniform as an increasing number of ionizing sources is visible to every point in the IGM.

Analytic models of the pre-overlap stage focus on the evolution of the H II filling factor, i.e., the fraction of the volume of the Universe which is filled by H II regions, $Q_{\rm H\ II}$. The modelling of individual H II regions can be used to understand the development of the total filling factor. Starting with equation (7.5), 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 $\bar{n}_\gamma$ equals the mean number of ionizing photons per baryon multiplied by the mean density of baryons $\bar{n}_b$. Following the arguments leading to equation (7.13), we find that if we include only stars

$$\frac{\bar{n}_\gamma}{\bar{n}_b} = N_{\rm ion} F_{\rm col} \ , \tag{7.15}$$

where the collapse fraction $F_{\rm col}$ is the fraction of all the baryons in the Universe which are in galaxies, i.e., the fraction of gas which has settled into halos and cooled efficiently inside them. In writing equation (7.15) we are assuming 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. The total number of ionizations equals the total number of ionizing photons produced by stars, i.e., all ionizing photons contribute regardless of the spatial distribution of sources. Also, the total recombination rate is proportional to the total ionized volume, regardless of its topology. Thus, even if two or more bubbles overlap, the model remains a good first approximation for $Q_{\rm H\ II}$ (at least until its value approaches unity).

Under these assumptions we convert equation (7.5), which describes individual H II regions, to an equation which statistically describes the transition from a neutral Universe to a fully ionized one:

$$\frac{dQ_{\rm H\ II}}{dt} = \frac{N_{\rm ion}}{0.76}\frac{dF_{\rm col}}{dt} - \alpha_B \frac{C}{a^3}\bar{n}_H^0 Q_{\rm H\ II} \ , \tag{7.16}$$

which admits the solution (in analogy with equation 7.7),

$$Q_{\rm H\ II}(t) = \int_0^t \frac{N_{\rm ion}}{0.76}\frac{dF_{\rm col}}{dt'}\, e^{F(t',t)} dt' \ , \tag{7.17}$$

where $F(t', t)$ is determined by equation (7.11).

A simple estimate of the collapse fraction at high redshift is the mass fraction (given by equation (3.35) 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 (3.30).

Although many models yield a reionization redshift around $z \sim 10$, the exact value depends on a number of uncertain parameters affecting both the source term and the recombination term in equation (7.16). The source parameters include the formation efficiency of stars and quasars and the escape fraction of ionizing photons produced by these sources.

The overlap of H II regions is expected to have occurred at different times in different regions of the IGM due to the cosmic scatter during the process of structure formation within finite spatial volumes. Reionization should be completed within a region of comoving radius R when the fraction of mass incorporated into collapsed objects in this region attains a certain critical value, corresponding to a threshold number of ionizing photons emitted per baryon. The ionization state of a region is governed by the factors of its enclosed ionizing luminosity, its over-density, and dense pockets of neutral gas within the region that are self shielding to ionizing radiation. 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 from the expression for the collapsed fraction $F_{\rm col}$ within a region of over-density $\bar{\delta}_{\rm R}$ on a comoving scale R, within the excursion-set formalism described in §3.3.1,

$$F_{\rm col} = {\rm erfc}\left[\frac{\delta_{\rm crit} - \bar{\delta}_{\rm R}}{\sqrt{2[\sigma^2_{\rm R_{min}} - \sigma^2_{\rm R}]}}\right] , \tag{7.18}$$

giving,[76]

$$\frac{\delta z}{(1+\bar{z})} = \frac{\bar{\delta}_{\rm R}}{\delta_{\rm crit}(\bar{z})} - \left[1 - \sqrt{1 - \frac{\sigma^2_{\rm R}}{\sigma^2_{\rm R_{min}}}}\right] , \tag{7.19}$$

where $\delta_{\rm crit}(\bar{z}) \propto (1+\bar{z})$ is the collapse threshold for an over-density at a redshift $\bar{z}$; and $\sigma_{\rm R}$ and $\sigma_{R_{\rm min}}$ are the variances in the power spectrum linearly extrapolated to $z = 0$ on comoving scales corresponding to the region of interest and to the minimum galaxy mass $M_{\rm min}$, respectively. The offset in the ionization redshift of a region depends on its linear over-density $\bar{\delta}_{\rm R}$. As a result, the distribution of offsets may be obtained directly from the power spectrum of primordial inhomogeneities. As can be seen from equation (7.19), larger regions have a smaller scatter due to their smaller cosmic variance. Interestingly, equation (7.19) is independent of the critical value of the collapsed fraction required for reionization.

The size distribution of ionized bubbles can also be calculated with an approximate analytic approach based on the excursion set formalism.[77] For a region to be ionized, galaxies inside it must produce a sufficient number of ionizing photons per baryon. This condition can be translated to the requirement that the collapsed fraction of mass in halos above some minimum mass $M_{\rm min}$ will exceed some threshold, namely $F_{\rm col} > \zeta^{-1}$. For example, requiring one ionizing photon per baryon would correspond to setting $\zeta = N_{\rm ion}$, where $N_{\rm ion}$ is defined in equation (7.12). We would like to find the largest region around every point whose collapse fraction satisfies the above condition on the collapse fraction, then calculate the abundance of ionized regions of this size. Different regions have different values of $F_{\rm col}$ because their mean density is different. Writing the collapse fraction in a region of mean overdensity δ_R as equation (7.18), we may derive the barrier condition on the mean overdensity within a region of mass $M = \frac{4\pi}{3} R^3 \rho_m$ in order for it to be ionized,

$$\delta_R > \delta_B(M,z) \equiv \delta_{\rm crit} - \sqrt{2} K(\zeta)[\sigma^2_{\rm min} - \sigma^2(M,z)]^{1/2}, \qquad (7.20)$$

where $K(\zeta) = {\rm erf}^{-1}(1-\zeta^{-1})$ and ${\rm erf}(x) \equiv 1 - {\rm erfc}(x)$. The barrier in equation (7.20) is well approximated by a linear dependence on σ^2 of the form, $\delta_B \approx B(M) = B_0 + B_1 \sigma^2(M)$, in which case the mass function has an analytic solution,

$$\frac{dn}{dM} = \sqrt{\frac{2}{\pi}} \frac{\rho_m}{M^2} \left| \frac{d\ln\sigma}{d\ln M} \right| \frac{B_0}{\sigma(M)} \exp\left[-\frac{B^2(M)}{2\sigma^2(M)} \right], \qquad (7.21)$$

where ρ_m is the mean comoving mass density. This solution for $(dn/dM)dM$ provides the comoving number density of ionized bubbles with IGM mass in the range between M and $M + dM$. The main difference between this result and the Press-Schechter mass function is that the barrier in this case becomes more difficult to cross on smaller scales because δ_B is a decreasing function of mass M. This gives bubbles a characteristic size. The size evolves with redshift in a way that depends only on ζ and $M_{\rm min}$.

A limitation of the above analytic model is that it ignores the non-local influence of sources on distant regions (such as voids) as well as the possible shadowing effect of intervening gas. Radiative transfer effects in the real Universe are inherently three-dimensional and cannot be fully captured by spherical averages as done in this model. Moreover, the value of $M_{\rm min}$ is expected to increase in regions that were already ionized, complicating the expectation of whether they will remain ionized later. Nevertheless, refined versions of this model agree well with rigorous radiative transfer simulations.[78]

7.3 SWISS CHEESE TOPOLOGY

Detailed numerical simulations which evolve the formation of galaxies along with the radiative transfer of the ionizing photons they produce within a representative cosmological volume, can provide a more accurate representation of the process of reionization than our approximate description above. The results from such simulations, illustrated in Figure 7.1, demonstrate that the spatial distribution of ionized bubbles is indeed determined by clustered groups of galaxies and not by individual galaxies. At early times, galaxies were strongly clustered even on very large scales (up to tens of Mpc), and these scales therefore dominate the structure of reionization. The basic idea is simple:[79] at high redshift, galactic halos are rare and correspond to high density peaks. As an analogy, imagine searching on Earth for mountain peaks above 5,000 meters. The 200 such peaks are not at all distributed uniformly, but instead are found in a few distinct clusters on top of large mountain ranges. Given the large-scale boost provided by a mountain range, a small-scale crest need only provide a small additional rise in order to become a 5,000 meter peak. The same crest, if it formed within a valley, would not come anywhere near 5,000 meters in total height. Similarly, in order to find the early galaxies, one should look in regions with large-scale density enhancements where galaxies are found in abundance.

The ionizing radiation emitted by the stars in each galaxy initially produces an isolated bubble of ionized gas. However, in a region dense with galaxies, the bubbles quickly overlap into one large bubble, completing reionization in this region even while the rest of the universe is still mostly neutral. Most importantly, since the abundance of rare density peaks is very sensitive to small changes in the density threshold, even a large-scale region with a small density enhancement (say, 10% above the mean density of the Universe) can have a much larger concentration of galaxies than in other regions (characterized, for example, by a 50% enhancement). On the other hand, reionization is more difficult to achieve in dense regions since the protons and electrons collide and recombine more often in such regions, and newly-formed hydrogen atoms need to be reionized again by additional ionizing photons. However, the overdense regions still end up reionizing first since the increase in the number of ionizing sources in these regions outweighs the higher recombination rate. The large-scale topology of reionization is therefore *inside out*, with underdense voids reionizing only at the very end of reionization using the help of extra ionizing photons coming in from their surroundings (which have a higher density of galaxies than the voids themselves). This is a key theoretical prediction awaiting observational testing.

Detailed analytical models accounting for large-scale variations in the abundance of galaxies confirm that the typical bubble size starts well below a Mpc early in reionization, as expected for an individual galaxy, rises to 5–10 Mpc during the central phase (i.e., when the Universe is half-ionized), and finally increases by another order of magnitude towards the end of reionization. These scales are given in comoving units that scale with the expansion of the universe such that the actual sizes at a redshift z were smaller than these numbers by a factor of $(1+z)$. Numerical simulations have only recently begun to reach the enormous scales needed to capture this evolution. Accounting precisely for gravitational evolution on a wide range of scales, but still crudely for gas dynamics, star formation, and the radiative transfer of ionizing photons, the simulations confirm that the large-scale topology of reionization is inside out, and that this “swiss cheese” topology can be used to study the abundance and clustering of the ionizing sources (Figures 7.1).

The characteristic observable size of the ionized bubbles at the end of reionization can be calculated based on simple considerations that depend only on the power spectrum of density fluctuations and the redshift. As the size of an ionized bubble increases, the time it takes a photon emitted by hydrogen to traverse it gets longer. At the same time, the variation in the time at which different regions reionize becomes smaller as the regions grow larger. Thus, there is a maximum region size above which the photon-crossing time is longer than the cosmic variance in reionization time. Regions larger than this size will be ionized at their near side

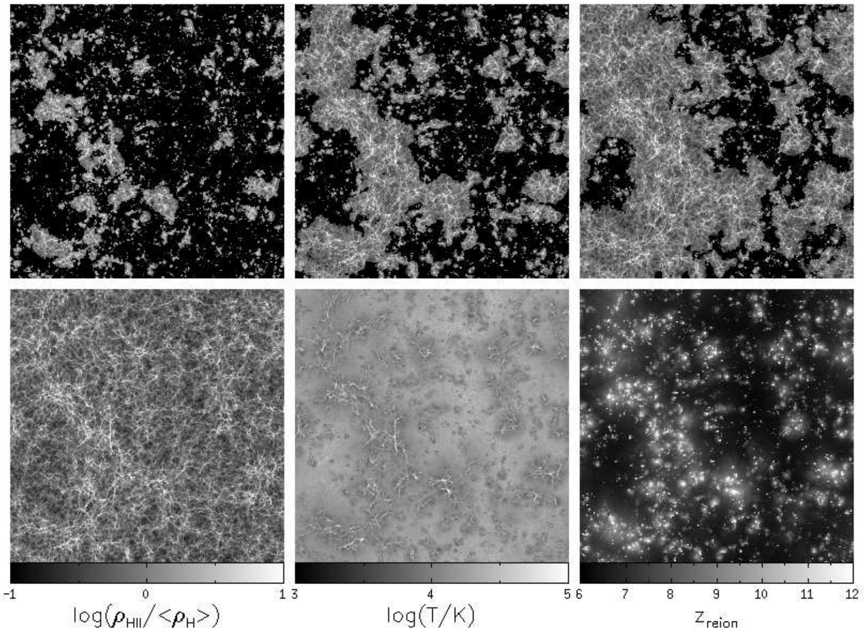

Figure 7.1 Snapshots from a numerical simulation illustrating the spatial structure of cosmic reionization in a slice of 140 comoving Mpc on a side. The simulation describes the dynamics of the dark matter and gas as well as the radiative transfer of ionizing radiation from galaxies. The first four panels (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. The resulting inhomogeneities in the temperature of the IGM introduce spatial variations in the cosmological Jeans mass, which in turn modulate the distribution of small galaxies (Babich, D., & Loeb, A. *Astrophys. J.* **640**, 1 (2006)) and the Lyman-α forest (Cen, R., McDonald, P., Trac, H., & Loeb, A. *Astrophys. J.*, submitted (2009)) at lower redshifts. Figure credit: Trac, H., Cen, R., & Loeb, A. *Astrophys. J.* **689**, L81 (2009).

by the time a photon crosses them towards the observer from their far side. They would appear to the observer as one-sided, and hence signal the end of reionization. Using $\sigma(M)$ in Figure 2.1, these considerations imply[80] a characteristic size for the ionized bubbles of ~ 10 physical Mpc at $z \sim 6$ (equivalent to 70 Mpc today). This result implies that future observations of the ionized bubbles (such as the low-frequency radio experiments described in §10) should be tuned to a characteristic angular scale of tens of arcminutes for an optimal detection of brightness fluctuations in the emission of hydrogen near the end of reionization.

7.4 REIONIZATION HISTORY

7.5 GLOBAL IONIZATION HISTORY

7.6 STATISTICAL DESCRIPTION OF SIZE DISTRIBUTION AND TOPOLOGY OF IONIZED REGIONS

7.7 RADIATIVE TRANSFER

7.8 RECOMBINATION OF IONIZED REGIONS

7.9 THE SOURCES OF REIONIZATION

7.9.1 Massive Stars

7.9.2 Quasars

7.9.3 Exotic Reionization Scenarios

7.10 HELIUM REIONIZATION

Chapter Eight

Feedback in the Early Universe

8.1 RADIATIVE FEEDBACK

8.1.1 Heating of the Intergalactic Medium: Mini-halos and the Clumping Factor

8.1.2 Photo-Heating and the Suppression of Low-Mass Galaxies

After the ionized bubbles overlapped in each region, the ionizing background increased sharply, and the IGM was heated by the ionizing radiation to a temperature $T_{\rm IGM} > 10^4$ K. Due to the substantial increase in the IGM pressure, the smallest mass scale into which the cosmic gas could fragment, the so-called Jeans mass, increased dramatically, changing the minimum mass of forming galaxies.

Gas infall depends sensitively on the Jeans mass. When a halo more massive than the Jeans mass begins to form, the gravity of its dark matter overcomes the gas pressure. Even in halos below the Jeans mass, although the gas is initially held up by pressure, once the dark matter collapses its increased gravity pulls in some gas. Thus, the Jeans mass is generally higher than the actual limiting mass for accretion. Before reionization, the IGM is cold and neutral, and the Jeans mass plays a secondary role in limiting galaxy formation compared to cooling. After reionization, the Jeans mass is increased by several orders of magnitude due to the photoionization heating of the IGM, and hence begins to play a dominant role in limiting the formation of stars. Gas infall in a reionized and heated Universe has been investigated in a number of numerical simulations. Three dimensional numerical simulations found a significant suppression of gas infall in even larger halos ($V_c \sim 75$ km s^{-1}), but this was mostly due to a suppression of late infall at $z < 2$.

When a volume of the IGM is ionized by stars, the gas is heated to a temperature $T_{\rm IGM} \sim 10^4$ K. If quasars dominate the UV background at reionization, their harder photon spectrum leads to $T_{\rm IGM} > 2 \times 10^4$ K. Including the effects of dark matter, a given temperature results in a linear Jeans mass corresponding to a halo circular velocity of

$$V_J \approx 80 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4 {\rm K}} \right)^{1/2} \ {\rm km\ s^{-1}}. \qquad (8.1)$$

In halos with a circular velocity well above V_J, the gas fraction in infalling gas equals the universal mean of Ω_b/Ω_m, but gas infall is suppressed in smaller halos. A simple estimate of the limiting circular velocity, below which halos have essentially no gas infall, is obtained by substituting the virial overdensity for the mean density in the definition of the Jeans mass. The resulting estimate is

$$V_{\rm lim} = 34 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4 {\rm K}} \right)^{1/2} \ {\rm km\,s^{-1}}. \tag{8.2}$$

This value is in rough agreement with the numerical simulations mentioned before.

Although the Jeans mass is closely related to the rate of gas infall at a given time, it does not directly yield the total gas residing in halos at a given time. The latter quantity depends on the entire history of gas accretion onto halos, as well as on the merger histories of halos, and an accurate description must involve a time-averaged Jeans mass. The gas content of halos in simulations is well fit by an expression which depends on the filtering mass, given by equation (3.12).

The reionization process was not perfectly synchronized throughout the Universe. Large-scale regions with a higher density than the mean tended to form galaxies first and reionized earlier than underdense regions. The suppression of low-mass galaxies by reionization is therefore modulated by the fluctuations in the timing of reionization. Inhomogeneous reionization imprint a signature on the power-spectrum of low-mass galaxies. Future high-redshift galaxy surveys hoping to constrain inflationary parameters must properly model the effects of reionization; conversely, they will also place new constraints on the thermal history of the IGM during reionization.

The increase in the minimum mass of star forming galaxies suppressed the global star formation rate per comoving volume after reionization.[81] In addition, the inhomogeneous nature of the reionization process modulated the distribution of the lowest-mass galaxies, and distorted their clustering properties relative to the underlying matter distribution.[82] The production and mixing of heavy elements was also modulated on the ~ 100 comoving Mpc scale of the largest H II regions.[83]

8.1.3 Recombination Radiation

8.2 LARGE-SCALE MECHANICAL FEEDBACK

8.3 CHEMICAL ENRICHMENT

Chapter Nine

The Lyman-α Line as a Probe of the Early Universe

9.1 HYDROGEN

Hydrogen is the most abundant element in the Universe. It is also the simplest atom possible, containing a proton and an electron held together by their mutual electric attraction. Because of its simplicity, the detailed understanding of the hydrogen atom structure played an important role in the development of quantum mechanics.

Since the lifetime of energy levels with principal quantum number n greater than 1 is far shorter than the typical time it takes to excite them in the rarefied environments of the Universe, hydrogen is commonly 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 9.1.

9.2 THE LYMAN-α TRANSITION

The most widely discussed transition of hydrogen in cosmology is the Lyman-α spectral line, 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[84] and, independently, Jim Gunn & Bruce Peterson[85] 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 $\sim 10^{-5}$. The lack of complete absorption for quasars at redshifts $z < 6.4$ is now interpreted as evidence that the diffuse IGM was fully ionized within less than a billion years after the Big Bang. Quasar spectra do show evidence for a so-called "forest" of Lyman-α absorption features, which originate from slight enhancements in the tiny fraction of hydrogen within overdense regions of the cosmic web. The Lyman-α forest has so far been observed in the spectrum of widely separated "skewers" pointing towards individual quasars. Since the cosmic web provides a measure of the power spectrum $P(k)$, there are plans to observe a dense array of skewers associated with a large number of quasars and map the related large-scale structure it delineates in three dimensions.

If a source were to be observed before or during the epoch of reionization, when the atomic fraction of hydrogen was more substantial, then all photons with wave-

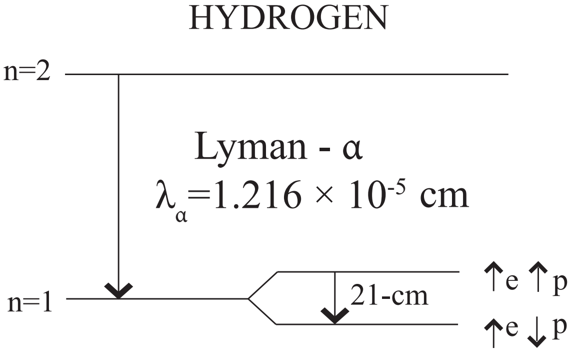

Figure 9.1 Two important transitions of the Hydrogen atom. The 21-cm transition of hydrogen is between two slightly separated (hyperfine) states of the ground energy level (principal quantum number $n = 1$). In the higher energy state, the spin of the electron (e) is aligned with that of the proton (p), 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).

lengths just short of the Lyman-α wavelength at the source (observed at $\lambda_\alpha = 1216(1 + z_s)$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. This would result in an observed absorption trough shortward of λ_α in the source spectrum, known as the "Gunn-Peterson effect". Often, quasars or gas-rich galaxies have a Lyman-α emission line, but the Gunn-Peterson effect is expected to eliminate the short-wavelength wing of the line (and potentially damp the entire Lyman-α emission feature if the line is sufficiently narrow).

The Gunn-Peterson trough serves as a robust indicator for the redshift of quasars, galaxies, and gamma-ray bursts during the epoch of reionization. Since it represents a broad spectral feature, its existence can be inferred by binning photons across broad frequency bands, a techniques labeled by astronomers as " photometry" (see Figure 9.2). This approach is particularly handy for faint galaxies that supply a small number of photons during an observing run, since fine binning of their frequency distribution (commonly termed as "spectroscopy") is impractical. In this context, the rarer bright sources have an important use. The Lyman-α cross-section is so large that absorption extends to wavelengths even slightly longer than λ_α, creating the appearance of a smooth wing with a characteristic shape. A spectroscopic detection of the detailed shape of this Lyman-α damping wing can be used to infer the neutral fraction of hydrogen in the IGM during reionization.[86]

The spectra of the highest-redshift quasars at $z < 6.4$ show hints of a Gunn-Peterson effect (see Figure 9.2), but the evidence is not conclusive since the observed high opacity of the Lyman-α transition can also result from trace amounts of hydrogen. A more fruitful approach would be to image hydrogen directly through

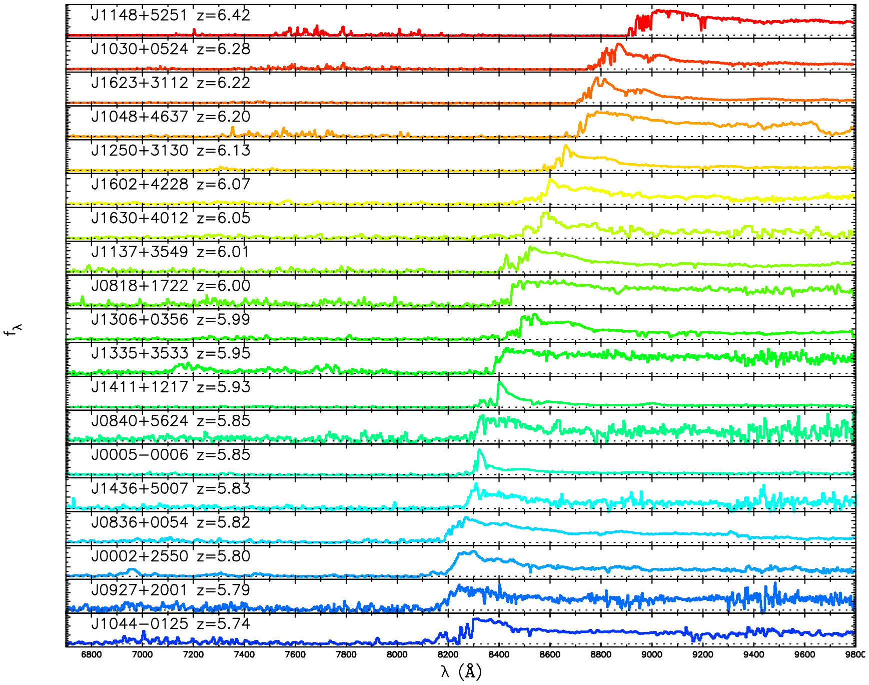

Figure 9.2 Observed spectra (flux per unit wavelength) of 19 quasars with redshifts $5.74 < z < 6.42$ from the Sloan Digital Sky Survey. For some of the highest-redshift quasars, the spectrum shows no transmitted flux shortward of the Lyman-α wavelength at the quasar redshift, providing a possible hint 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).

a ground-state transition with a weaker opacity, a possibility we explore next.

9.3 LYMAN-α EMISSION FROM GALAXIES

9.4 LYMAN-α SCATTERING IN THE DIFFUSE IGM

9.4.1 Basic Principles

The scattering cross-section of the Lyman-α resonance line by neutral hydrogen is given by

$$\sigma_\alpha(\nu) = \frac{3\lambda_\alpha^2\Lambda_\alpha^2}{8\pi}\frac{(\nu/\nu_\alpha)^4}{4\pi^2(\nu-\nu_\alpha)^2+(\Lambda_\alpha^2/4)(\nu/\nu_\alpha)^6}, \tag{9.1}$$

where $\Lambda_\alpha = (8\pi^2 e^2 f_\alpha/3m_e c\lambda_\alpha^2) = 6.25\times10^8\ \mathrm{s}^{-1}$ is the Lyman-α ($2p \rightarrow 1s$) decay rate, $f_\alpha = 0.4162$ is the oscillator strength, and $\lambda_\alpha = 1216\text{Å}$ and $\nu_\alpha = (c/\lambda_\alpha) = 2.47\times10^{15}$ Hz are the wavelength and frequency of the Lyman-α line. The term in the numerator is responsible for the classical Rayleigh scattering.

We consider a source at a redshift z_s beyond the redshift of reionization[i] $z_{\rm reion}$, and the corresponding scattering optical depth of a uniform, neutral IGM of hydrogen density $n_{\rm H,0}(1+z)^3$ between the source and the reionization redshift. The optical depth is a function of the observed wavelength $\lambda_{\rm obs}$,

$$\tau(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz\frac{cdt}{dz}n_{\rm H,0}(1+z)^3\sigma_\alpha\left[\nu_{\rm obs}(1+z)\right], \tag{9.2}$$

where $\nu_{\rm obs} = c/\lambda_{\rm obs}$ and

$$\frac{dt}{dz} = -\left[(1+z)H(z)\right]^{-1} = -H_0^{-1}\left[\Omega_m(1+z)^5+\Omega_\Lambda(1+z)^2\right]^{-1/2}. \tag{9.3}$$

At wavelengths longer than Lyman-α at the source, the optical depth obtains a small value; these photons redshift away from the line center along its red wing and never resonate with the line core on their way to the observer. Considering only the regime in which $|\nu-\nu_\alpha| \gg \Lambda_\alpha$, we may ignore the second term in the denominator of equation (9.1). This leads to an analytical result for the red damping wing of the Gunn-Peterson trough,

$$\tau(\lambda_{\rm obs}) = \tau_s\left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)\tilde{\lambda}_{\rm obs}^{3/2}\left[I(\tilde{\lambda}_{\rm obs}^{-1}) - I([(1+z_{\rm reion})/(1+z_s)]\tilde{\lambda}_{\rm obs}^{-1})\right], \tag{9.4}$$

for $\tilde{\lambda}_{\rm obs} \geq 1$, where we define

$$\tilde{\lambda}_{\rm obs} \equiv \frac{\lambda_{\rm obs}}{(1+z_s)\lambda_\alpha} \tag{9.5}$$

[i] We define the reionization redshift to be the redshift at which the individual H II regions overlapped and most of the IGM volume was ionized. In most realistic scenarios, this transition occurs rapidly on a time-scale much shorter than the age of the universe. This is mainly due to the short distances between neighboring sources.

and

$$I(x) \equiv \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\ln\left[\frac{1+x^{1/2}}{1-x^{1/2}}\right] . \quad (9.6)$$

The total optical depth is obtained from the integral (10.4) by substituting $dz = [(1+z)/\nu_\alpha]d\nu$, by taking out of the integral all components evaluated at the resonance redshift except the sharply-peaked $\sigma_\alpha(\nu)$, and by using equation (6.35). This gives,

$$\tau_s = \frac{\pi e^2 f_\alpha \lambda_\alpha n_{\rm H\,I}\,(z_s)}{m_e c H(z_s)} \approx 6.45 \times 10^5 x_{\rm H\,I} \left(\frac{\Omega_b h}{0.03}\right)\left(\frac{\Omega_m}{0.3}\right)^{-1/2}\left(\frac{1+z_s}{10}\right)^{3/2} . \quad (9.7)$$

Here, $H \approx 100h\,\Omega_m^{1/2}(1+z_s)^{3/2}\ {\rm km\,s^{-1}\ Mpc^{-1}}$ is the Hubble parameter at the source redshift $z_s >> 1$, $f_\alpha = 0.4162$ and $\lambda_\alpha = 1216$Å are the oscillator strength and the wavelength of the Lyman-α transition; $n_{\rm H\,I}\,(z_s)$ is the neutral hydrogen density at the source redshift (assuming primordial abundances); Ω_m and Ω_b are the present-day density parameters of all matter and of baryons, respectively; and $x_{\rm H\,I}$ is the average fraction of neutral hydrogen.

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 (9.7). Since $\tau_s \sim 10^5$, the flux from the source is entirely suppressed in this regime. Similarly, the Lyman-β resonance produces another trough at wavelengths $(1+z_{\rm reion})\lambda_\beta \leq \lambda \leq (1+z_s)\lambda_\beta$, 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 in-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 are absorbed when they photoionize atoms of hydrogen or helium. 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,

$$\sigma_{bf}(\nu) = \frac{6.30 \times 10^{-18}}{Z^2}\ {\rm cm}^2 \times \left(\frac{\nu_0}{\nu}\right)^4 \frac{e^{4-(4\tan^{-1}\epsilon)/\epsilon}}{1-e^{-2\pi/\epsilon}} \quad \text{for } \nu \geq \nu_0, \quad (9.8)$$

where

$$\epsilon \equiv \sqrt{\frac{\nu}{\nu_0} - 1}. \quad (9.9)$$

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). 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,[87]

$$\sigma_{bf,{\rm He\ I}}(\nu) = 9.492 \times 10^{-16}\ {\rm cm}^2 \times \left[(x-1)^2 + 4.158\right] \times y^{-1.953}\left(1 + 0.825y^{1/4}\right)^{-3.188} , \quad (9.10)$$

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

For rough estimates, the average photoionization cross-section for a mixture of hydrogen and helium with cosmic abundances can be approximated in the range of $54 < h\nu < 10^3$ eV as $\sigma_{bf} \approx \sigma_0(\nu/\nu_{\mathrm{H},0})^{-3}$, where $\sigma_0 \approx 6 \times 10^{-17}\ \mathrm{cm}^2$. 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, $t(z_{\mathrm{reion}}) - t(z_s)$. At high redshifts this yields,

$$\begin{aligned}\tau_{bf}(\lambda_{\mathrm{obs}}) &= \int_{z_{\mathrm{reion}}}^{z_s} dz \frac{cdt}{dz} n_0(1+z)^3 \sigma_{\mathrm{bf}}\left[\nu_{\mathrm{obs}}(1+z)\right] \\ &\approx 1.5 \times 10^2 \left(\frac{\lambda}{100\text{Å}}\right)^3 \left[\frac{1}{(1+z_{\mathrm{reion}})^{3/2}} - \frac{1}{(1+z_s)^{3/2}}\right]. \quad (9.11)\end{aligned}$$

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.

The transmitted flux between the Gunn-Peterson troughs due to Lyman-α and Lyman-β absorption is suppressed by the Lyman-α forest in the post-reionization epoch. Transmission of flux due to ionized bubbles in the pre-reionization epoch is expected to be negligible. The redshift of reionization can be inferred in principle from the spectral shape of the red damping wing or from the transmitted flux between the Lyman series lines. However, these signatures are complicated in reality by damped Lyman-α systems along the line of sight or by the inhomogeneity or peculiar velocity field of the IGM in the vicinity of the source. Moreover, bright sources, such as quasars, tend to ionize their surrounding environment and the resulting H II region in the IGM could shift the Lyman-α trough substantially.

The inference of the Lyman-α transmission properties of the IGM from the observed spectrum of high-redshift sources suffers from uncertainties about the precise emission spectrum of these sources, and in particular the shape of their Lyman-α emission line. The first galaxies and quasars are expected to have pronounced recombination lines of hydrogen and helium due to the lack of dust in their interstellar medium. Lines such as Hα or the He II 1640 Å line should reach the observer unaffected by the intervening IGM, since their wavelength is longer than that of the Lyman-α transition which dominates the IGM opacity. However, as described above, the situation is different for the Lyman-α line photons from the source. As long as $z_s > z_{\mathrm{reion}}$, the intervening neutral IGM acts like a fog and obscures the view of the Lyman-α line itself [in contrast to the situation with sources at $z_s < z_{\mathrm{reion}}$, where most of the intervening IGM is ionized and only the flux of photons with wavelengths shorter than Lyman-α is suppressed by the Lyman-α forest]. Photons which are emitted at the Lyman-α line center have an initial scattering optical depth of $\sim 10^5$ in the surrounding medium.

The Lyman-α line photons are not destroyed but instead are absorbed and re-emitted[ii]. Due to the Hubble expansion of the IGM around the source, the frequency

[ii] At the redshifts of interest, $z_s \sim 10$, the low densities and lack of chemical enrichment of the IGM make the destruction of Lyman-α photons by two-photon decay or dust absorption unimportant.

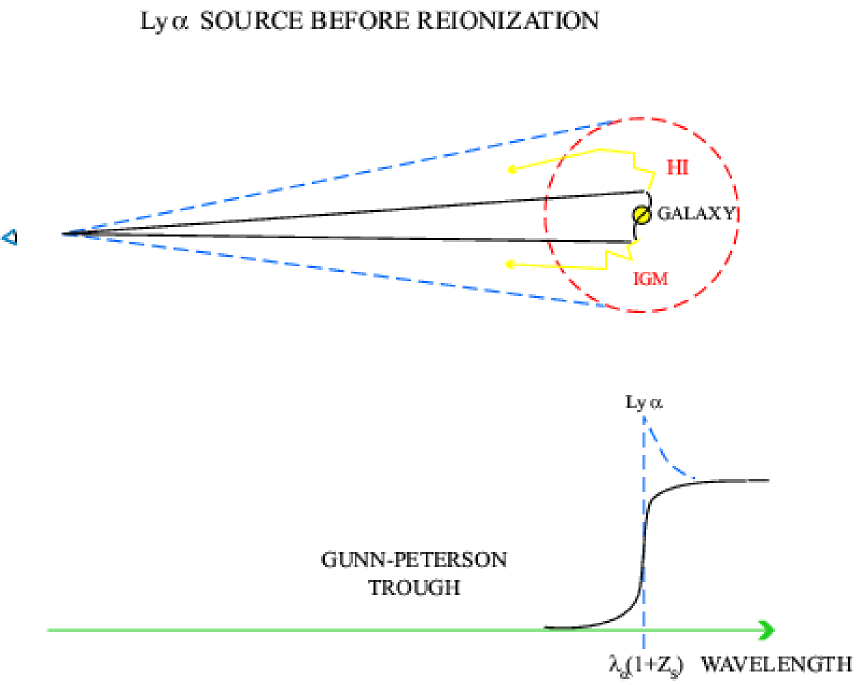

Figure 9.3 Halo of scattered Lyman-α line photons from a galaxy embedded in the neutral IGM prior to reionization (also called *Loeb-Rybicki halo*). The line photons diffuse in frequency due to the Hubble expansion of the surrounding medium and eventually redshift out of resonance and escape to infinity. 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.

of the photons is slightly shifted by the Doppler effect in each scattering event. As a result, the photons diffuse in frequency to the red side of the Lyman-α resonance. Eventually, when their net frequency redshift is sufficiently large, they escape and travel freely towards the observer (see Figure 9.3). As a result, the source creates a faint Lyman-α halo on the sky[iii]. The well-defined radiative transfer problem of a point source of Lyman-α photons embedded in a uniform, expanding neutral IGM was solved by Loeb & Rybicki (1999).[88] The Lyman-α halo 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 [equation (9.4)] equal to unity. At high redshifts, the leading term in equation (9.4) yields

$$\nu_\star = 8.85 \times 10^{12}\ \mathrm{Hz} \times \left(\frac{\Omega_b h}{0.05\sqrt{\Omega_m}}\right)\left(\frac{1+z_s}{10}\right)^{3/2}, \tag{9.12}$$

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,

$$r_\star = 1.1 \left(\frac{\Omega_b/0.05}{\Omega_m/0.3}\right)\ \mathrm{Mpc}. \tag{9.13}$$

Typically, the 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 line profile as shown in Figures 9.3 and 9.4. The scattered photons are highly polarized and so the shape of the halo would be different if viewed through a polarization filter.

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. On sufficiently large scales where the Hubble flow is smooth and the gas is neutral, the Lyman-α brightness distribution can be used to determine the cosmological mass densities of baryons and matter. 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 at somewhat longer wavelengths. The disappearance of Lyman-α halos below a certain redshift can be used to determine z_reion.

[iii] The photons 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. The Gunn-Peterson halo is made even dimmer relative to the Lyman-α line halo by the fact that the luminosity of the source per unit frequency is often much lower in the continuum than in the Lyman-α line.

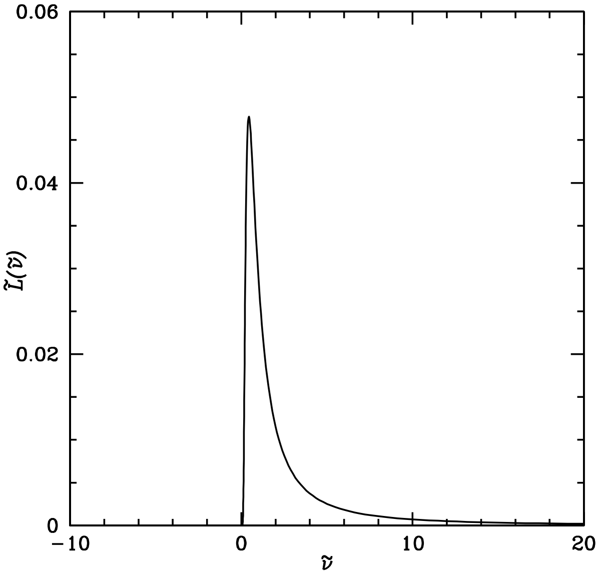

Figure 9.4 Monochromatic photon luminosity of a Lyman-α halo as a function of normalized frequency shift from the Lyman-α resonance, $\tilde{\nu} \equiv (\nu_\alpha - \nu)/\nu_\star$. The observed spectral flux of photons $F(\nu)$ (in photons cm^{-2} s^{-1} Hz^{-1}) from the entire Lyman-α halo is $F(\nu) = (\tilde{L}(\tilde{\nu})/4\pi d_{\rm L}^2)(\dot{N}_\alpha/\nu_\star)(1+z_s)^2$ where $\dot{N}_\alpha$ is the production rate of Lyman-α photons by the source (in $\mathrm{photons\,s^{-1}}$), $\nu = \tilde{\nu}\nu_\star/(1+z_{\rm s})$, and $d_{\rm L}$ is the luminosity distance to the source [from Loeb, A. & Rybicki, G. B. *Astrophys. J.* **524**, 527 (1999); see also **520**, L79 (1999)].

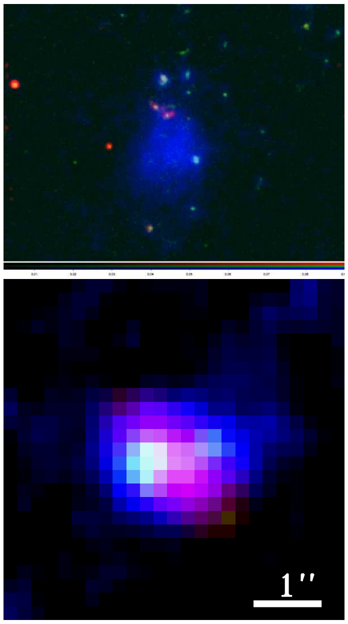

Figure 9.5 *Top:* A false color image of a Lyman-α blob (LAB) at a redshift $z = 2.656$. The hydrogen Lyman-α emission is shown in blue, and images in the optical V-band and the near-infrared J and H bands are shown in green and red, respectively. Note the compact galaxies lying near the northern (top) 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).

9.4.2 Lyman-α Blobs

Since their initial discovery,[89] several tens of Lyman-α blobs (LABs) have been found[90] in the redshift range $z \sim$ 2–7. 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[91] of $\sim 10^3 M_\odot$ yr^{-1}, or with obscured active galactic nuclei (AGN).[92] However, other blobs have been found that are not associated with any source powerful enough to explain the observed Lyman-α luminosities.[93] The observed LABs have a large spatial extent, ~ 150 kpc, and a diffuse elliptical or filamentary morphology. Blobs were observed at redshifts as high as[94] $z = 6.6$, as illustrated in Fig. 9.5.

The origin of LABs is still being debated. Some models relate LABs to cooling radiation from gas assembling into the cores of galaxies.[95] Other models invoke photoionization of cold ($T \sim 10^4$ K), dense, spatially extended gas by an obscured AGN[96] or extended X-Ray emission;[97] the compression of ambient gas by superwinds to a dense Lyman-α emitting shell;[98] or star formation triggered by relativistic jets from AGN.[99] The latest models[100] relate LABs to filamentary flows of cold ($\sim 10^4$K) gas into galaxies, which are generically found in numerical simulations of galaxy formation.[101] These cold flows contain ~ 5–15% of the total gas content[102] in halos as massive as $M_{\rm halo} \sim 10^{12}$–$10^{13}M_{\odot}$.

Chapter Ten

The 21-cm Line

10.1 ATMOIC PHYSICS

The radiative transfer equation for a spectral line reads,

$$\frac{dI_\nu}{d\ell} = \frac{\phi(\nu)h\nu}{4\pi}\left[n_2 A_{21} - \left(n_1 B_{12} - n_2 B_{21}\right) I_\nu\right], \tag{10.1}$$

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), subscripts 1 and 2 denote the lower and upper 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. We can then make use of the standard relations in atomic physics: $B_{21} = (g_1/g_2)B_{12}$ and $B_{21} = A_{21}(c^2/2h\nu^3)$, where g is the spin degeneracy factor of each state. For the 21cm transition, $A_{21} = 2.85 \times 10^{-15}\mathrm{s}^{-1}$ 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,

$$\left(\frac{n_2}{n_1}\right) = \left(\frac{g_2}{g_1}\right)\exp\left\{-\frac{E}{k_B T_s}\right\}, \tag{10.2}$$

where $E/k_B = 68$ mK is the transition energy. In the regime of interest, E/k 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. We may also replace ΔI_ν with $2k_B T_b/\lambda^2$, since the frequencies of interest are on the Rayleigh-Jeans wing of the CMB blackbody spectrum. By substituting all the above relations in equation (10.1), we get the observed deviation from the CMB brightness temperature T_b in terms of the optical depth in the 21cm line $\tau \ll 1$,

$$T_b = \frac{\tau}{(1+z)}\left(T_s - T_\gamma\right). \tag{10.3}$$

In an expanding Universe with a uniform hydrogen number density $n_{\rm HI}$ and with a velocity gradient equal to the Hubble parameter H, the 21-cm optical depth can be derived similarly to equation(9.7). Writing $\phi(\nu) \sim 1/(\Delta\nu)$ we get $\Delta I_\nu \propto \Delta\ell\phi(\nu)\nu = |cdt/dz|(\nu dz/d\nu) = c/H$, giving

$$\tau = \frac{3}{32\pi}\frac{h^3 c^3 A_{21}}{E^2}\frac{n_{\rm HI}}{H k_B T_s}. \tag{10.4}$$

In the presence of inhomogeneities, $n_{\rm HI}$ should include the density fluctuations and H should be replaced by the full velocity gradient, which includes the line-of-sight derivative of the line-of-sight component of the peculiar velocity. The latter introduces an anisotropy to the power spectrum of 21-cm brightness fluctuations.

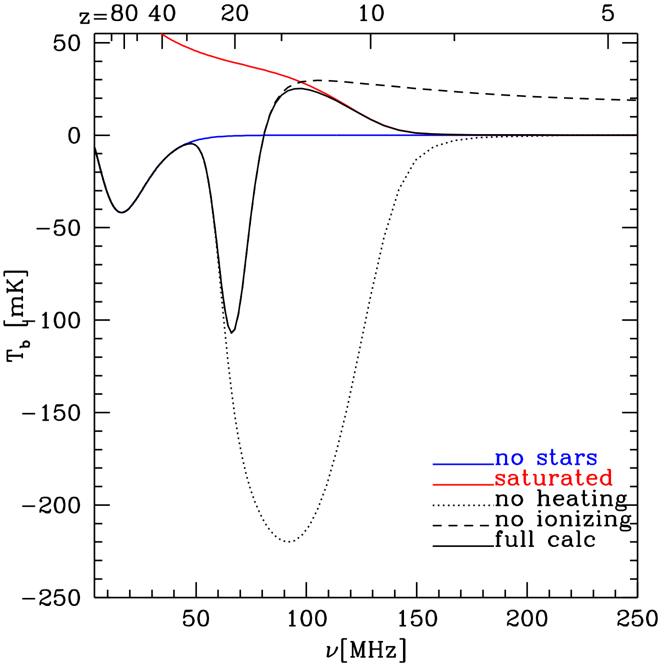

Figure 10.1 Evolution of the global (spectral) 21-cm signal for different scenarios. *Solid blue curve:* no stars; *solid red curve:* $T_S \gg T_\gamma$; *black dotted curve:* no heating; *black dashed curve:* no ionization; *black solid curve:* full calculation. **Figure credit:** Pritchard, J., & Loeb, A. Phys. Rev. **D**, in press (2010).

The right-hand-side of equation (10.3) is proportional to $(T_s - T_\gamma)$, yielding an emission signal if $T_s > T_\gamma$ and an absorption signal otherwise. In addition, equations (10.3) and (10.4) imply that as long as $T_s \gg T_{\rm cmb}$, the magnitude of the emission signal T_b is not dependent on the existence of the CMB (or even the particular value of T_s, as $\tau \propto T_s^{-1}$).

10.2 INTERACTION WITH GAS AND UV/X-RAY RADIATION BACKGROUNDS

10.3 STATISTICAL AND IMAGING TOOLS

10.4 OBSERVATIONAL PROSPECTS

In difference from interferometric arrays, single dipole experiments which integrate over most of the sky, can search for the global (spectral) 21-cm signal shown in Figure 10.1. Examples of such experiments are CoRE or EDGES (*http://www.haystack.mit.edu/ast/arrays/Edges/*). Rapid reionization histories which span a redshift range $\Delta z < 2$ can be constrained, provided that local foregrounds (see Figure 10.2) can be well modelled by low-order polynomials in frequency. Observations in the frequency range 50-100 MHz can potentially constrain the Lyman-α and X-ray emissivity of the first stars forming at redshifts $z \sim 15$–25, as illustrated in Figure 10.3.

10.5 THE TRANSITION TO THE POST-REIONIZATION UNIVERSE

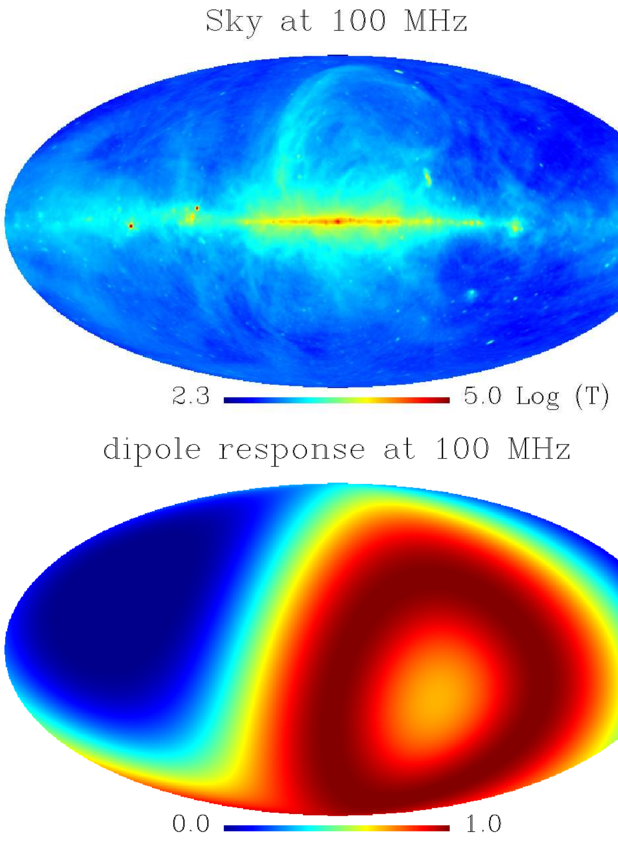

Figure 10.2 *Top panel:* Radio map of the sky at 100 MHz. *Bottom panel:* Ideal dipole response averaged over 24 hours. **Figure credits:** Pritchard, J., & Loeb, A. Phys. Rev. **D**, in press (2010); de Oliveira-Costa, A. et al. Mon. Not. R. Astron. Soc. **388**, 247 (2008).

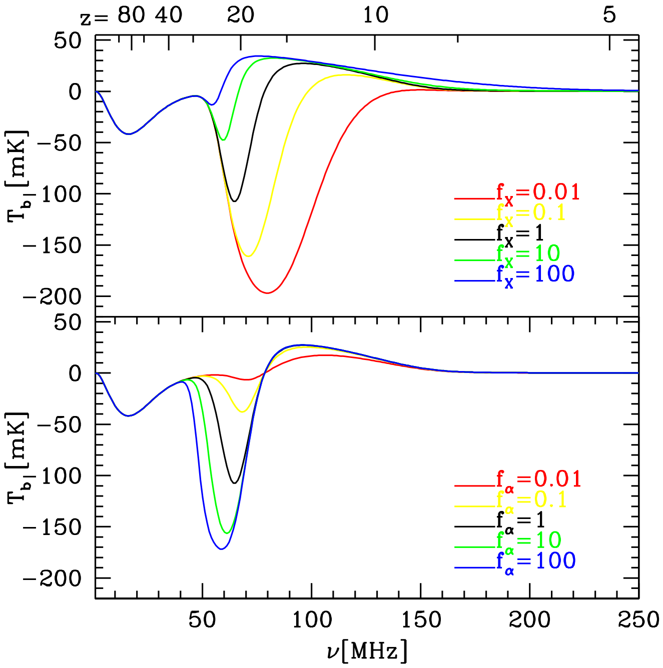

Figure 10.3 Dependence of global 21-cm signal on the X-ray (top panel) and Lyman-α (bottom panel) emissivity of stars. Each case depicts examples with the characteristic emissivity reduced or increased by a factor of up to 100. **Figure credit:** Pritchard, J., & Loeb, A. Phys. Rev. **D**, in press (2010).

Chapter Eleven

The 21-cm Line: Additional Notes (by A.L.)

In quantum mechanics, elementary particles possess a fundamental property called "spin" (classically thought of as the rotation of a particle around its axis), which has a half-integer or integer magnitude, and an up or down state. The ground state of hydrogen is split into two very close ("hyperfine") states: an upper energy level (triplet state) in which the spin of the electron is lined up with that of the proton, and a lower energy level (singlet state) in which the two are anti-aligned. The transition to the lower level is accompanied by the emission of a photon with a wavelength of 21 centimeter (see Figure 9.1). The 21-cm transition was theoretically predicted by Hendrick van de Hulst in 1944 and detected in 1951 by Harold Ewen & Ed Purcell, who put a horn antenna out of an office window in the Harvard physics department and saw the 21-cm emission from the Milky Way.

The existence of neutral hydrogen prior to reionization offers the prospect of detecting its 21-cm emission or absorption relative to the CMB.[103] The optical depth is only $\sim 1\%$ in this case for a fully neutral IGM, making the 21-cm line a more suitable probe for the epoch of reionization than the Lyman-α line. By observing different wavelengths of 21cm$\times(1+z)$, one is slicing the Universe at different redshifts z. 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. A full map of the distribution of atomic hydrogen (denoted by astronomers as H I) as a function of redshift would provide a three-dimensional image of the swiss-cheese structure of the IGM during reionization, as illustrated in image 11.1. The cavities in the hydrogen distribution are the ionized bubbles around groups of early galaxies.

The relative population of the two levels defines the so-called *spin (excitation) temperature*, which may deviate from the ordinary (kinetic) temperature of the gas in the presence of a radiation field. The coupling between the gas and the microwave background (owing to the small residual fraction of free electrons left over from the hydrogen formation epoch) kept the gas temperature equal to the radiation temperature up to 10 million years after the Big Bang. Subsequently, the cosmic expansion cooled the gas faster than the radiation, and collisions among the atoms maintained their spin temperature at equilibrium with their own kinetic temperature. At this phase, cosmic hydrogen can be detected in *absorption* against the microwave background sky since it is colder. Regions that are somewhat denser than the mean will produce more absorption and underdense regions will produce less absorption. The resulting fluctuations in the 21-cm brightness simply reflects the primordial inhomogeneities in the gas.[104] A hundred million years after the Big Bang, cosmic expansion diluted the density of the gas to the point where the

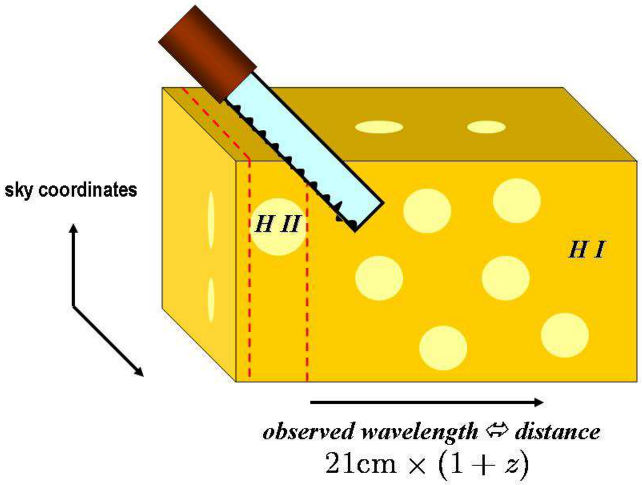

Figure 11.1 21-cm imaging of ionized bubbles during the epoch of reionization is analogous to slicing swiss cheese. The technique of slicing at intervals separated by the typical dimension of a bubble is optimal for revealing different patterns in each slice.

collisional coupling of the spin temperature to the gas became weaker than its coupling to the microwave background. At this stage, the spin temperature returned to equilibrium with the radiation temperature, and it became impossible to see the gas against the microwave background brightness. Once the first galaxies lit up, they heated the gas (mainly by emitting X-rays which penetrated the thick column of intergalactic hydrogen) as well as its spin temperature (through UV photons which couple the spin temperature to the gas kinetic temperature). The increase of the spin temperature beyond the microwave background temperature requires much less energy per atom than ionization, so this heating occurred well before the Universe was reionized. Once the spin temperature had risen above the microwave background (CMB) temperature, the gas could be seen against the microwave sky in *emission*. At this stage, the hydrogen distribution is punctuated with bubbles of ionized gas which are created around groups of galaxies. Below we describe these evolutionary stages more quantitatively.

The basic physics of the hydrogen spin transition is determined as follows. The ground-state hyperfine levels of hydrogen tend to thermalize with the CMB, making the IGM unobservable. If other processes shift the hyperfine level populations away from thermal equilibrium, then the gas becomes observable against the CMB either in emission or in absorption. The relative occupancy of the spin levels is usually described in terms of the hydrogen spin temperature T_s, defined by

$$\frac{n_1}{n_0} = 3 \exp\left\{-\frac{T_*}{T_s}\right\} , \tag{11.1}$$

where n_0 and n_1 refer respectively to the singlet and triplet hyperfine levels in the atomic ground state ($n = 1$), and $T_* = 0.068$ K is defined by $k_B T_* = E_{21}$, where the energy of the 21 cm transition is $E_{21} = 5.9 \times 10^{-6}$ eV, corresponding to a photon frequency of 1420 MHz. In the presence of the CMB alone, the spin states reach thermal equilibrium with $T_s = T_\gamma = 2.73(1+z)$ K on a time-scale of $\sim T_*/(T_\gamma A_{10}) = 3 \times 10^5 (1+z)^{-1}$ yr, where $A_{10} = 2.87 \times 10^{-15}$ 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.

The IGM is observable only when the kinetic temperature T_k of the gas (defined by the motion of its atoms) differs from T_γ, and an effective mechanism couples T_s to T_k. At early times, collisions dominate this coupling because the gas density is still high, but once a significant galaxy population forms in the Universe, the spin temperature is affected also by an indirect mechanism that acts through the scattering of Lyman-α photons, the so-called Wouthuysen-Field effect, named after the Dutch physicist Siegfried Wouthuysen and Harvard astrophysicist George Field who explored it first.[105] Here continuum UV photons produced by early radiation sources redshift by the Hubble expansion into the local Lyman-α line at a lower redshift and mix the spin states.

A patch of neutral hydrogen at the mean density and with a uniform T_s produces (after correcting for stimulated emission) an optical depth at an observed wavelength of $21(1+z)$ cm of

$$\tau(z) = 1.1 \times 10^{-2} \left(\frac{T_\gamma}{T_s}\right) \left(\frac{1+z}{10}\right)^{1/2} , \tag{11.2}$$

where we have assumed $z \gg 1$. The observed spectral intensity I_ν relative to the CMB at a frequency ν is measured by radio astronomers as an effective brightness temperature T_b of blackbody emission at this frequency, defined using the Rayleigh-Jeans limit of the Planck formula, $I_\nu \equiv 2k_B T_b \nu^2/c^2$.

The brightness temperature through the IGM is $T_b = T_\gamma e^{-\tau} + T_s(1 - e^{-\tau})$, so the observed differential antenna temperature of this region relative to the CMB is[106]

$$\begin{aligned} T_b &= (1+z)^{-1}(T_s - T_\gamma)(1 - e^{-\tau}) \\ &\simeq 29\,\mathrm{mK} \left(\frac{1+z}{10}\right)^{1/2} \left(\frac{T_s - T_\gamma}{T_s}\right) , \end{aligned} \tag{11.3}$$

where we have made use of the fact that $\tau \ll 1$ and T_b has been redshifted to $z = 0$. The abbreviated unit 'mK' stands for milli-degree K or 10^{-3}K.

In overdense regions, the observed T_b is proportional to the overdensity, while in partially ionized regions T_b is proportional to the neutral fraction. Also, if $T_s \gg T_\gamma$, then the IGM is observed in emission at a level that is independent of T_s. On the other hand, if $T_s \ll T_\gamma$, then the IGM is observed in absorption[107] at a level enhanced by a factor of T_γ/T_s. As a result, a number of cosmic events are expected to leave observable signatures in the redshifted 21-cm line, as discussed below.

Figure 11.2 illustrates the mean IGM evolution for three examples in which reionization is completed at different redshifts, $z = 6.47$ (thin lines), $z = 9.76$ (medium thickness lines), and $z = 11.76$ (thick lines). The top panel shows the

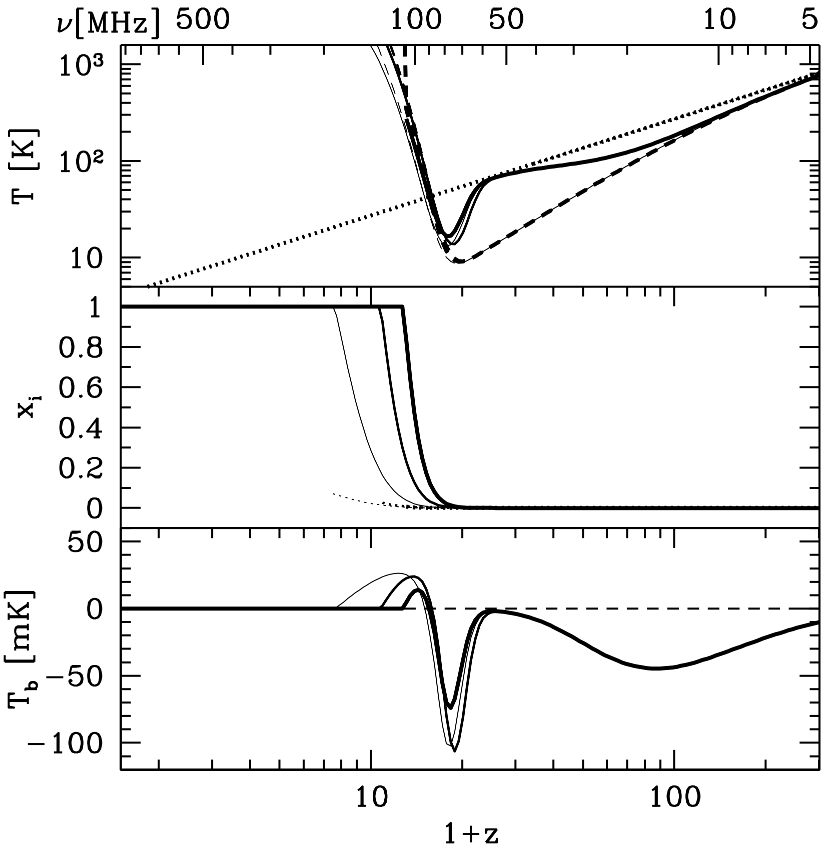

Figure 11.2 *Top panel:* Evolution with redshift z of the CMB temperature $T_{\rm CMB}$ (dotted curve), gas kinetic temperature T_k (dashed curve), and spin temperature T_s (solid curve). Following cosmological recombination at a redshift $z \sim 10^3$, the gas temperature tracks the CMB temperature ($\propto (1+z)$) down to a redshift $z \sim 200$ and then declines below it ($\propto (1+z)^2$), until the first X-ray sources (accreting black holes or exploding supernovae) heat it up well above the CMB temperature. The spin temperature of the 21-cm transition interpolates between the gas and CMB temperatures; initially it tracks the gas temperature through atomic collisions, then it tracks the CMB through radiative coupling, and eventually it tracks the gas temperature once again after the production of a cosmic background of UV photons that redshift into the Lyman-α resonance. *Middle panel:* Evolution of the gas fraction in ionized regions x_i (solid curve) and the ionized fraction outside these regions (due to diffuse X-rays) x_e (dotted curve). *Bottom panel:* Evolution of mean 21-cm brightness temperature T_b. The horizontal axis at the top provides the observed photon frequency for the different redshifts shown at the bottom. Each panel shows curves for three models in which reionization is completed at different redshifts: $z = 6.47$ (thin lines), $z = 9.76$ (medium thickness lines), and $z = 11.76$ (thick lines). Figure credit: Pritchard, J. & Loeb, A. *Phys. Rev.* **D78**, 103511 (2008).

global evolution of the CMB temperature $T_{\rm CMB}$ (dotted curve), the gas kinetic temperature T_k (dashed curve), and the spin temperature T_s (solid curve). The middle panel shows the evolution of the ionized gas fraction and the bottom panel displays the mean 21 cm brightness temperature, T_b.

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. These teams plan to assemble arrays of thousands of dipole antennae and correlate their electric field measurements. Although the radio technology for the frequency range of interest is the same as used in past decades for TV or radio communication, the experiments have never been done before because computers were not sufficiently powerful to analyze and correlate the large volume of data produced by these arrays. The planned experiments include the Low Frequency Array,[108] the Murchison Wide-Field Array shown in Figure 11.3,[109] the Primeval Structure Telescope,[110] the Precision Array to Probe the Epoch of Reionization,[111] and ultimately the Square Kilometer Array.[112] These low-frequency radio observatories will search over the next decade for redshifted 21-cm emission or absorption from redshifts $z \sim 6.5$–15, corresponding to observed wavelengths of 1.5–3.4 meters (comparable to the height of a person). Current observational projects in 21-cm cosmology are at the same status as CMB research was prior to the first statistical detection of the sky temperature fluctuations by the COsmic Background Explorer (COBE) satellite.

Because the smallest angular size that can be resolved by a telescope is of order the observed wavelength divided by the telescope diameter, radio astronomy at wavelengths as large as a meter has remained relatively undeveloped. Producing resolved images even of large sources such as cosmological ionized bubbles requires telescopes which have a kilometer scale. It is much more cost-effective to use a large array of thousands of simple antennas distributed over several kilometers, and to use computers to cross-correlate the measurements of the individual antennae and combine them effectively into a single large telescope. The new experiments are located mostly in remote sites, because the frequency band of interest overlaps with more mundane terrestrial telecommunications.[i]

Detection of the redshifted 21-cm signal is challenging. Relativistic electrons within our Milky Way galaxy produce synchrotron radio emission as they gyrate around the Galactic magnetic field.[ii] This results in a radio foreground that is larger than the expected reionization signal by at least a factor of ten thousand. But not all is lost. By shifting slightly in observed wavelength one is slicing the hydrogen distribution at different redshifts and hence one is seeing a different map of its bubble structure, but the synchrotron foreground remains nearly the same. Theoretical calculations demonstrate that it is possible to extract the signal from the epoch of

[i] These experiments will bring a new capability to search for leakage of TV/radio signals from a distant civilization (Loeb, A., & Zaldarriaga, M. *J. of Cosm. and Astro-Part. Phys.* **1**, 20 (2007)). Post World-War II leakage of radio signals from our civilization could have been detected by the same experiments out to a distance of tens of light years. Since our civilization produced its brightest radio signals for military purposes during the cold war, it is plausible that the brightest civilizations out there are the most militant ones. Therefore, if we do detect a signal, we better not respond.

[ii] For a pedagogical description of synchrotron radiation, see §6 in Rybicki, G. B., & Lightman, A. P. *Radiative Processes in Astrophysics*, Wiley, New York (1979).

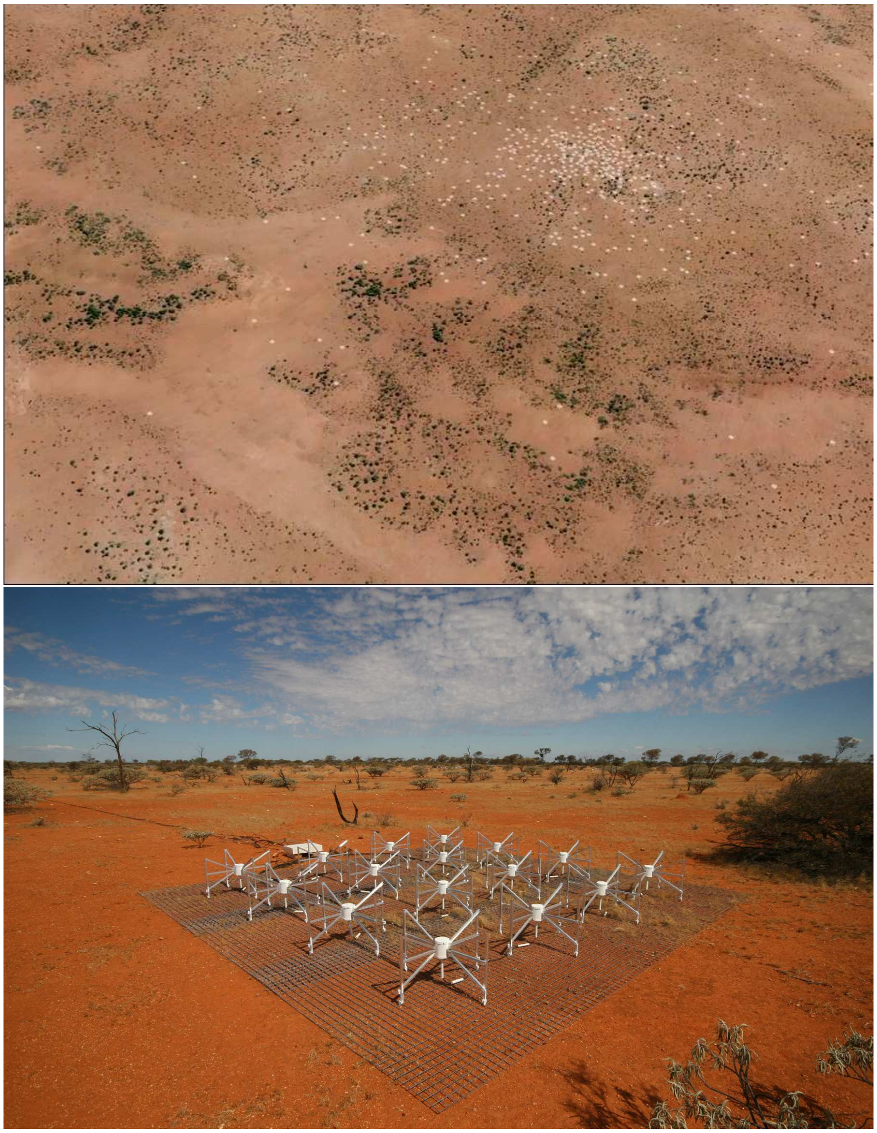

Figure 11.3 *Top:* Artificial illustration of the expected MWA experiment with 512 tiles (white spots) of 16 dipole antennae each, spread across an area of 1.5 km in diameter in the desert of western Australia. With a collecting area of 8,000 square meters, the array will be sensitive to 21-cm emission from cosmic hydrogen in the redshift range of z =6–15 by operating in the radio frequency range of 80–300 MHz. *Bottom:* An actual image of one of the tiles. Image credits: Bowman, J. & Lonsdale, C. (2009).

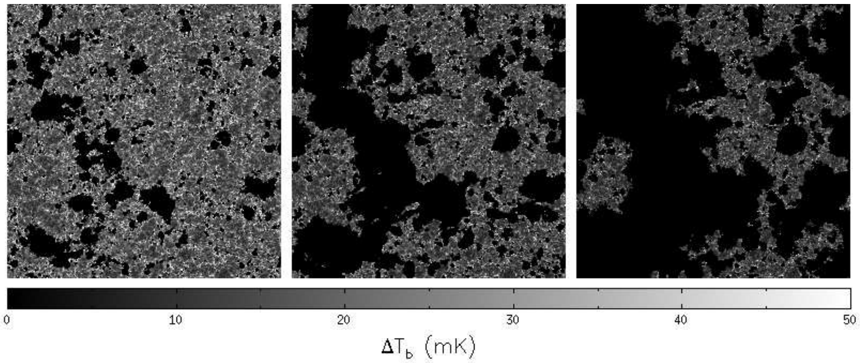

Figure 11.4 Map of the fluctuations in the 21 cm brightness temperature on the sky, ΔT_b (mK), based on a numerical simulation which follows the dynamics of dark matter and gas in the IGM as well as the radiative transfer of ionizing photons from galaxies. The panels show the evolution of the signal in a slice of 140 comoving Mpc on a side, in three snapshots corresponding to the simulated volume being 25, 50, and 75 % ionized. These snapshots correspond to the top three panels in Figure 7.1. Since neutral regions correspond to strong emission (i.e., a high T_b), the 21-cm maps illustrate the global progress of reionization and the substantial large-scale spatial fluctuations in the reionization history. Figure credit: Trac, H., Cen, R., & Loeb, A. *Astrophys. J.* **689**, L81 (2009).

reionization by subtracting the radio images of the sky at slightly different wavelengths. In approaching redshifted 21-cm observations, although the first inkling might be to consider the mean emission signal in the bottom panel of Figure 11.2 (and this is indeed the goal of some single-antenna experiments[113]), the signal is orders of magnitude fainter than the synchrotron foreground (see Figure 11.5). Thus, most observers have focused on the expected characteristic variations in T_b, both with position on the sky and especially with frequency, which signifies redshift for the cosmic signal. The synchrotron foreground is expected to have a smooth frequency spectrum, so it is possible to isolate the cosmological signal by taking the difference in the sky brightness fluctuations at slightly different frequencies (as long as the frequency separation corresponds to the characteristic size of ionized bubbles). Large-scale patterns in the 21-cm brightness from reionization are driven by spatial variations in the abundance of galaxies; the 21-cm fluctuations reach a root-mean-square amplitude of $\sim$5 mK in brightness temperature on a scale of 10 comoving Mpc (Figure 11.4). While detailed maps will be difficult to extract due to the foreground emission, a statistical detection of these fluctuations (through the power spectrum) is expected to be well within the capabilities of the first-generation experiments now being built. Current work suggests that the key information on the topology and timing of reionization can be extracted statistically.[114]

While numerical simulations of reionization are now reaching the cosmological box sizes needed to predict the large-scale topology of the ionized bubbles, they often do so at the price of limited small-scale resolution. These simulations cannot yet follow in any detail the formation of individual stars within galaxies, or the feedback from stars on the surrounding gas, such as heating by radiation or the piston effect of supernova explosions, which blow hot bubbles of gas enriched with the chemical products of stellar nucleosynthesis. The simulations cannot directly predict whether the stars that form during reionization are similar to the stars in the Milky Way and nearby galaxies or to the primordial $100M_{\odot}$ stars. They also cannot determine whether feedback prevents low-mass dark matter halos from forming stars. Thus, models are needed to make it possible to vary all these astrophysical parameters of the ionizing sources and study the effect they have on the 21-cm observations.

The current theoretical expectations for reionization and for the 21-cm signal are based on rather large extrapolations from observed galaxies to deduce the properties of much smaller galaxies that formed at an earlier cosmic epoch. Considerable surprises are thus possible, such as an early population of quasars, or even unstable exotic particles that emitted ionizing radiation as they decayed. The forthcoming observational data in 21-cm cosmology should make the next decade a very exciting time.

It is of particular interest to separate signatures of the fundamental physics, such as the initial conditions from inflation and the nature of the dark matter and dark energy, from the astrophysics, involving phenomena related to star formation, which cannot be modeled accurately from first principles. This is particularly easy to do before the first galaxies formed ($z > 25$), at which time the 21-cm fluctuations are expected to simply trace the primordial power spectrum of matter density perturbations which is shaped by the initial conditions from inflation and the dark matter. The same simplicity applies after reionization ($z < 6$) – when only dense pockets of self-shielded hydrogen (associated with individual galaxies) survive, and those behave as test particles and simply trace the matter distribution.[115] During the epoch of reionization, however, the 21-cm fluctuations are mainly shaped by the topology of ionized regions, and thus depend on uncertain astrophysical details involving star formation. However, even during this epoch, the imprint of deviations from the Hubble flow (i.e., peculiar velocity fluctuations $\mathbf{u}$ which are induced gravitationally by density fluctuations δ; see equations 2.3-2.4), can in principle be used to separate the signatures of fundamental physics from the astrophysics.

Deviations from the smooth Hubble flow imprint a particular form of anisotropy in the 21-cm fluctuations caused by gas motions along the line of sight. This anisotropy, expected in any measurement of density based on a resonance line (or on any other redshift indicator), results from velocity compression. Consider a photon traveling along the line of sight that resonates with absorbing atoms at a particular point. In a uniform, expanding universe, the absorption optical depth encountered by this photon probes only a narrow strip of atoms, since the expansion of the universe makes all other neighboring atoms move with a relative velocity which takes them outside the narrow frequency width of the resonance line. If, however, there is a density peak near the resonating position, the increased gravity will reduce the expansion velocities around this point and bring more gas into the resonating velocity width. The associated Doppler effect is sensitive only to the line of sight component of the velocity gradient of the gas, and thus causes an observed anisotropy in the power spectrum even when all physical causes of the fluctuations

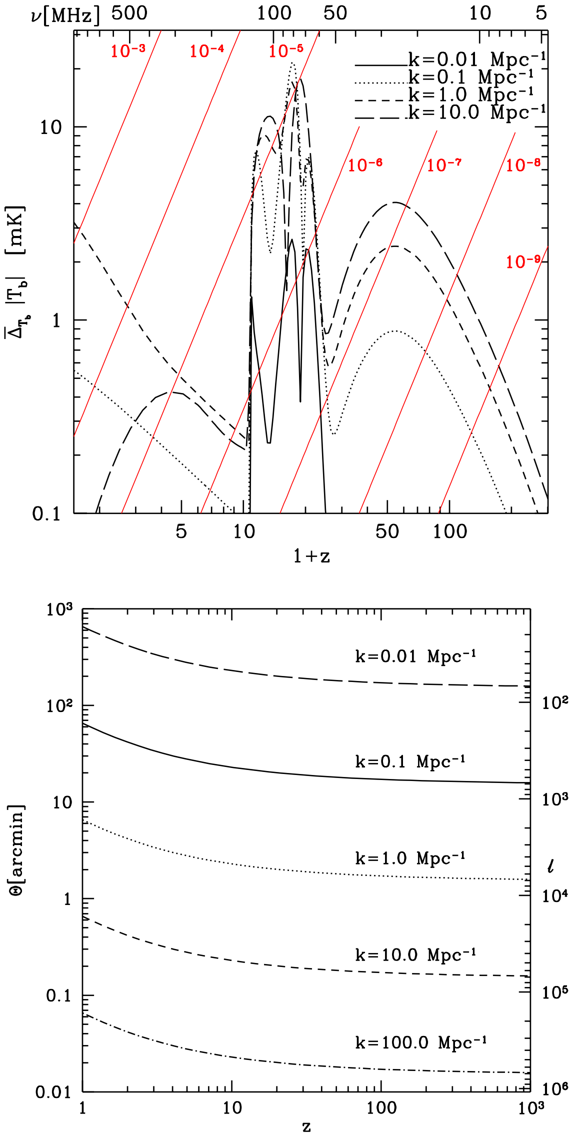

Figure 11.5 *Top:* Predicted redshift evolution of the angle-averaged amplitude of the 21-cm power spectrum ($|\bar{\Delta}_{T_b}| = [k^3 P_{21-\rm cm}(k)/2\pi^2]^{1/2}$) at comoving wavenumbers $k = 0.01$ (solid curve), 0.1 (dotted curve), 1.0 (short dashed curve), 10.0 (long dashed curve), and $100.0\rm Mpc^{-1}$ (dot-dashed curve). In the model shown, reionization is completed at $z = 9.76$. The horizontal axis at the top shows the observed photon frequency at the different redshifts. The diagonal straight lines show various factors of suppression for the synchrotron Galactic foreground, necessary to reveal the 21-cm signal. *Bottom:* Redshift evolution of the angular scale on the sky corresponding different comoving wavenumbers, $\Theta = (2\pi/k)/d_{\rm A}$. The labels on the right-hand side map angles to angular moments (often used to denote the multipole index of a spherical harmonics decomposition of the sky), using the approximate relation $\ell \approx \pi/\Theta$. Along the line of sight (the third dimension), an observed frequency bandwidth $\Delta\nu$ corresponds to a comoving distance of $\sim 1.8\ {\rm Mpc}(\Delta\nu/0.1\ {\rm MHz})[(1+z)/10]^{1/2}$. Figure credit: Pritchard, J. & Loeb, A. *Phys. Rev.* **D78**, 103511 (2008).

are statistically isotropic. This anisotropy implies that the observed power spectrum in redshift space $P(\mathbf{k})$ depends on the angle between the line of sight and the $\mathbf{k}$-vector of a Fourier mode, not only on its amplitude k. This angular dependence allows to separate out the simple gravitational signature of density perturbations from the complex astrophysical effects of reionization, such as star and black hole formation, and feedback from supernovae and quasars.[116]

11.1 OBSERVING MOST OF THE OBSERVABLE VOLUME

In general, cosmological surveys are able to measure the spatial power spectrum of primordial density fluctuations, $P(\mathbf{k})$, to a precision that is ultimately limited by Poisson statistics of the number of independent regions (the so-called "cosmic variance"). The 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 $z \sim 10^3$. 21-cm observations are advantageous because they access a three-dimensional volume instead of the two-dimensional surface probed by the CMB, and hence cover a larger number of independent regions in which the primordial initial conditions were realized. Moreover, the expected 21-cm power extends down to the pressure-dominated (Jeans) scale of the cosmic gas which is orders of magnitude smaller than the comoving scale at which the CMB anisotropies are damped by photon diffusion. Consequently, the 21-cm photons can trace the primordial inhomogeneities with a much finer resolution (i.e. many more independent pixels) than the CMB. Also, 21-cm studies promise to extend to much higher redshifts than existing galaxy surveys, thereby covering a much bigger fraction of the comoving volume of the observable Universe. At these high redshifts, small-scale modes are still in the perturbative (linear growth) regime where their statistical analysis is straightforward (§2.1). Altogether, the above factors imply that the 21-cm mapping of cosmic hydrogen may potentially carry the largest number of bits of information about the initial conditions of our Universe compared to any other survey method in cosmology.[117]

The limitations of existing data sets (on which the cosmological parameters in Table 1.1 are based) are apparent in Figure 1.3, which illustrates the comoving volume of the Universe out to a redshift z as a function of z. State-of-the-art galaxy redshift surveys, such as the spectroscopic sample of luminous red galaxies (LRGs) in the first Sloan Digital Sky Survey (SDSS), extended only out to $z \sim 0.3$ (only one tenth of our horizon) and probed $\sim 0.1\%$ of the observable comoving volume of the Universe. Surveys of the 21-cm emission (or a large number of quasar skewers through the Lyman-α forest) promise to open a new window into the distribution of matter through the remaining 99.9% of the cosmic volume. These ambitious experiments might also probe the gravitational growth of structure through most of the observable universe, and provide a new test of Einstein's theory of gravity across large scales of space and time.

Chapter Twelve

Observations of High-Redshift Galaxies and Implications for Reionization

12.1 GENERAL BACKGROUND

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 simoultaneously. 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 quickly 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 bubbles overlapped and the initially neutral gas in between the galaxies was completely reionized.

Although the above progression of events sounds plausible, at this time it is only a thought floating in the minds of theorists that has not yet received confirmation from observational data. Empirical cosmologists would like to actually see direct evidence for the reionization process before accepting it as common knowledge. *How can one observe the reionization history of the Universe directly?*

One way is to search for the radiation emitted by the first galaxies using large new telescopes from the ground as well as from space. Another way is to image hydrogen and study the cavities of ionized bubbles within it. The observational exploration of the reionization epoch promises to be one of the most active frontiers in cosmology over the coming decade.

12.1.1 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 $10^{19}\ \mathrm{cm} = 3$pc 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 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,

$$f = \frac{L}{4\pi d_{\mathrm{L}}^2}, \tag{12.1}$$

where d_{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_{em} and is

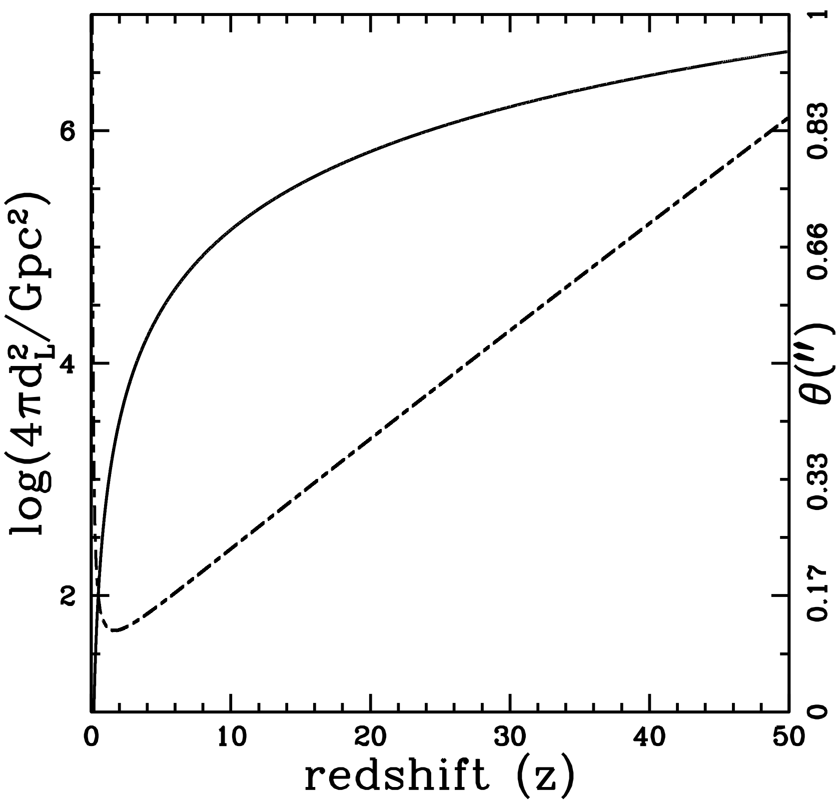

Figure 12.1 The solid line (corresponding to the label on the left-hand side) shows Log$_{10}$ of the conversion factor between the luminosity of a source and its observed flux, $4\pi d_{\rm L}^2$ (in Gpc^2), as a function of redshift, z. The dashed-dotted line (labeled on the right) gives the angle θ (in arcseconds) occupied by a galaxy of a 1 kpc diameter as a function of redshift.

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:

$$r_{\rm em} = \int_{t_{\rm em}}^{t_{\rm obs}} \frac{cdt}{a(t)} = \frac{c}{H_0}\int_0^z \frac{dz'}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}, \tag{12.2}$$

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[i] 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

$$f = \frac{Ldt_{\rm em}/(1+z)}{4\pi r_{\rm em}^2 dt_{\rm obs}} = \frac{L}{4\pi r_{\rm em}^2(1+z)^2}, \tag{12.3}$$

implying that[ii]

$$d_{\rm L} = r_{\rm em}(1+z) = d_{\rm A}(1+z)^2. \tag{12.4}$$

The area dilution factor $4\pi d_{\rm L}^2$ is plotted as a function of redshift in the bottom panel of Figure 12.10. 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 12.9 compares the predicted flux per unit frequency[iii] from a galaxy at a redshift $z_s = 10$ for a Salpeter IMF and for massive ($> 100M_\odot$) 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 3.2 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.[118] 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.[119]

[i] 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.

[ii] 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 .

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

12.1.2 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.[iv] 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 12.2). 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.[v] The HDF contained many red galaxies with some reaching a redshift as high as 6, or even higher.[120] 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.[vi] Red galaxies were identified in the HUDF image up to a redshift of $z \sim 7$, and possibly even higher, showing that the typical UV luminosity of galaxies declines with redshift at $z > 4$.[121] The redshifts of galaxies are inferred either through a search for a Lyman-α emission line (identifying so-called "Lyman-α galaxies"),[122] or through a search for a spectral break associated with the absorption of intervening hydrogen (so-called "Lyman-break galaxies").[123] For very faint sources, redshifts are only identified crudely based on the spectral trough produced by hydrogen absorption in the host galaxy and the IGM (see §7).

The abundance of Lyman-α galaxies shows a strong decline between $z = 5.7$ and $z = 7$, as expected from a correspondingly rapid increase in the neutral fraction of the IGM (which would scatter the Lyman-α line photons and make the line emission from these galaxies undetectable),[124] but this interpretation is not unique.

[iv] Turner, E. private communication (2009).

[v] Just prior to the HDF, an important paper about high-redshift galaxies was declined for publication because the referee pointed out the "well-known fact" that there are no galaxies beyond a redshift of 1.

[vi] In order for galaxy surveys to be statistically reliable, they need to cover large areas of the sky. Counts of galaxies in small fields of view suffer from a large cosmic variance owing to galaxy clustering.

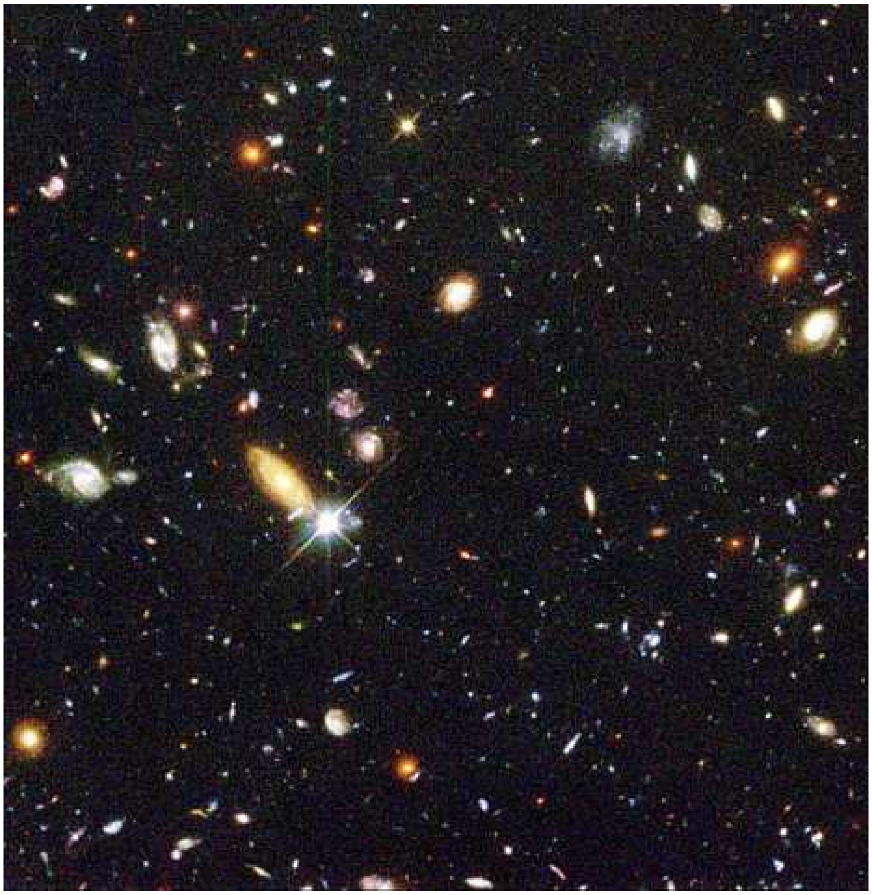

Figure 12.2 The first Hubble Deep Field (HDF) image taken in 1995. The HDF covers an area 2.5 arcminute across and contains a few thousand galaxies (with a few candidates up to a redshift $z \sim 6$). The image was taken in four broadband filters centered on wavelengths of 3000, 4500, 6060, and 8140Å, with an average exposure time of ~ 0.127 million seconds per filter.

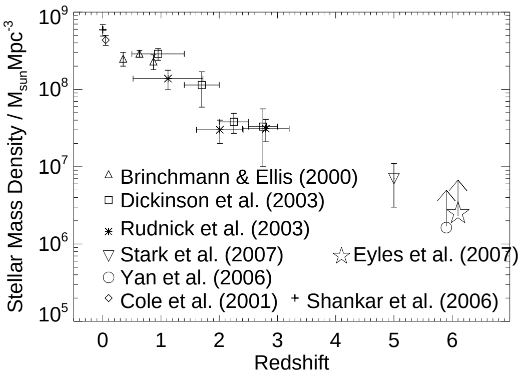

Figure 12.3 The observed evolution in the cosmic mass density of stars (in solar masses per comoving Mpc^3) as a function of redshift [data compiled by Eyles, L. P., et al. *Mon. Not. R. Astron. Soc.* **374**, 910 (2007); for additional recent data, see Labbe, I., et al. *Astrophys. J.* **708**, L26 (2010).]. The estimates at $z > 5$ should be regarded as lower limits due to the missing contribution of low-luminosity galaxies below the detection threshold [Stark, D., et al. *Astrophys. J.* **659**, 84 (2007)]. Nevertheless, the data shows that less than a few percent of all present-day stars had formed at $z > 5$, in the first 1.2 billion years after the Big Bang. A minimum density $\sim 1.7 \times 10^6 f_{\rm esc}^{-1} \rm M_\odot\ Mpc^{-3}$ of Population II 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 in the Universe.

The mass budget of stars at $z \sim 5$–6 has been inferred from complementary infrared observation with the Spitzer Space Telescope (see Figure 12.3). The mean age of the stars in individual galaxies implies that they had formed at $z \sim 10$ and could have produced sufficient photons to reionize the IGM. The advantage of measuring the cumulative stellar density rather than the star formation rate density (see Figure 12.12) is that the cumulative density provides a census of stars that were made in faint galaxies below the detection threshold, that were incorporated at a later time into detectable galaxies. This method would particularly effective with JWST.

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[125] 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.[126]

So far, we have not seen the first generation of dwarf galaxies at redshifts $z > 10$ that were responsible for reionization.

12.1.3 Observing the First Gamma-Ray Bursts

Explosions of individual massive stars (such as supernovae) can also 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 §5.5). These afterglows can be used to study the first stars directly. Also, similarly to quasars, these beacons of light probe the state of the cosmic gas through its absorption line signatures on their spectra along the line of sight. GRBs were discovered by the *Swift* satellite out to a record redshift of $z = 8.3$, merely 620 million years after the Big Bang, and significantly earlier than the farthest known quasar ($z = 6.4$, see Figure 9.2). It is already evident that GRB observations hold the promise of opening a new window into the infant Universe.

Standard light bulbs appear fainter with increasing redshift, but this is not the case with GRBs which are transient events that fade with time. When observing a burst at a constant 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.[127] 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 12.4.

In contrast to bright quasars, 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. 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.[128] Unfortunately, the ability of GRBs to probe the neutral fraction of the IGM is very often compromised by damped Lyman-α absorption of hydrogen within their host galaxy.

As discussed in §5.5, long-duration GRBs are believed to originate from the collapse of massive stars at the end of their lives (Figure 5.8). Since the very first stars were likely massive, they could have produced GRBs.[129] If they did, 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 photomoetric 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.

GRBs are expected to trace the star formation history better than galaxy surveys

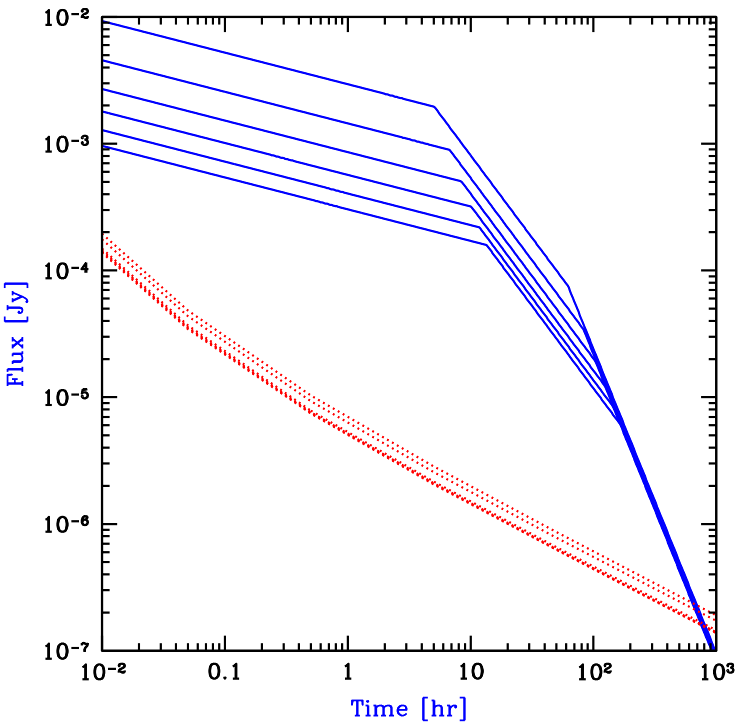

Figure 12.4 Detectability of high-redshift GRB afterglows as a function of time since the GRB explosion as measured by the observer. The GRB afterglow flux (in Jy) is shown at the redshifted Lyman-α wavelength (solid curves). Also shown (dotted curves) is a crude estimate for the spectroscopic detection 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).

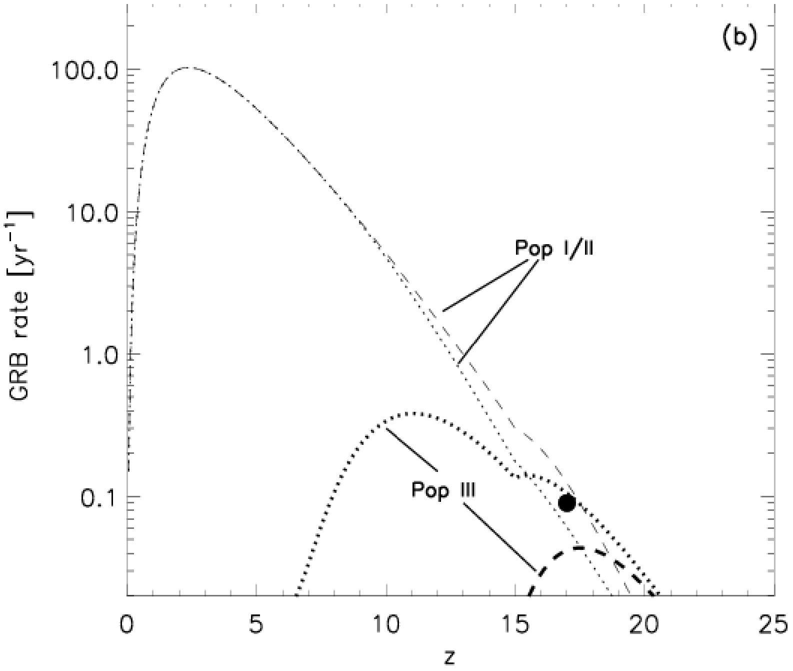

Figure 12.5 Theoretically predicted rate of observable GRBs as a function of redshift (assuming that the GRB rate is proportional to the cosmic star formation rate at all redshifts). *Dotted lines:* Contribution to the observed GRB rate from Pop I/II and Pop III for the case of slow metal enrichment of the IGM. *Dashed lines:* Contribution to the GRB rate from Pop I/II and Pop III for the case of rapid metal enrichment of the IGM. *Filled circle:* GRB rate from Pop III stars if these, in an extreme scenario, were responsible for reionizing the universe at $z \sim 17$. Figure credit: Bromm, V., & Loeb, A. *Astrophys. J.* **642**, 382 (2006).

since they flag the typical sites of star formation irrespective of how faint their host galaxy is. The survey method selects galaxies above some observed flux threshold and miss stars in low-luminosity galaxies. This introduces an artificial suppression (whose magnitude increases with redshift) in the inferred cosmic star formation rate per comoving volume. Even though the observed GRB rate does not suffer from this redshift-dependent suppression, it is naturally expected to decline sharply with increasing redshift as illustrated in Figure 12.5.

12.1.4 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

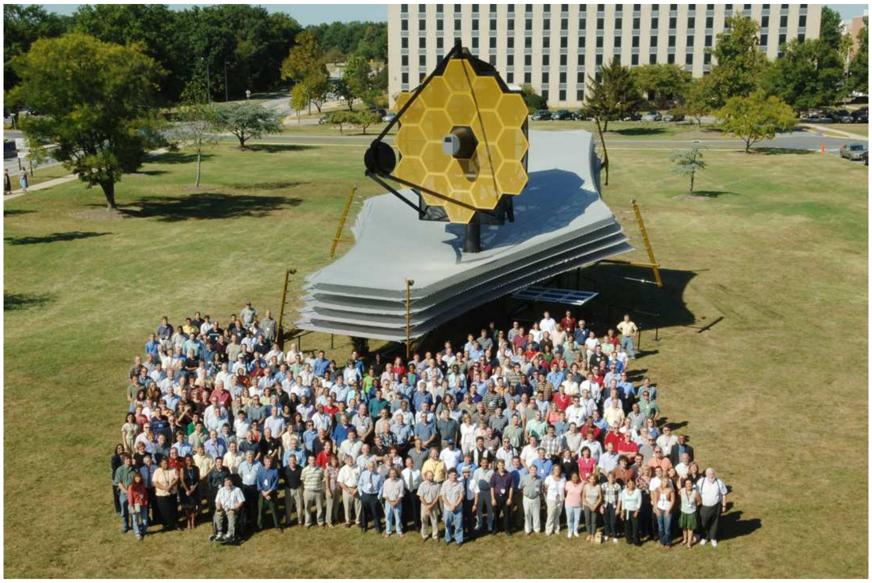

Figure 12.6 A full scale model of the James Webb Space Telescope (JWST), the successor to the Hubble Space Telescope (http://www.jwst.nasa.gov/). JWST includes a primary mirror 6.5 meters in diameter, and offers instrument sensitivity across the infrared wavelength range of 0.6–28μm which will allow 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.

infrared wavelength range of 0.6–28μm (see illustration 12.6). 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. The earliest galaxies are expected to be extremely faint and compact, for two reasons: first, they are associated with the smallest gaseous objects to have condensed out of the primordial gas, and second they are located at the greatest distances from us among all galaxies.[130] Endowed with a large aperture and positioned outside the Earth's atmospheric emissions and opacity, JWST is ideally suited for resolving the faint glow from the first galaxies. It would be particularly exciting if JWST finds spectroscopic evidence for metal-free (Population III) stars. As shown in Figure 12.9, the smoking gun signature would be a spectrum with no metal lines, a strong UV continuum consistent with a blackbody spectrum of $\sim 10^5$ K truncated by an IGM absorption trough (at wavelengths shorter than Lyman-α in the source frame; see §7.2), and showing strong helium recombination lines, including a line at 1640Å to which the IGM is transparent, from the interstellar gas around these hot stars.[131]

Several initiatives to construct large infrared telescopes on the ground are also underway. The next generation of ground-based telescopes will have an effective diameter of 24-42 meters; examples include the European Extremely Large Telescope,[132] the Giant Magellan Telescope,[133] and the Thirty Meter Telescope,[134] which are illustrated in Figure 12.7. Along with JWST, they will be able to image and survey a large sample of early galaxies. 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[135] (see §10). 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).[136]

What makes the study of the first galaxies so exciting is that it involves work in progress. If all the problems were solved, there would be nothing left to be discovered by future scientists, such as some of the young readers of this book. 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 derive the implications of the observed properties of high-redshift galaxies to reionization.

12.2 MASS FUNCTION OF STARS

Present-day stars were traditionally classified into two two populations. Population I stars like the Sun are luminous, hot, metal-rich, young stars commonly found in the disks of spiral galaxies. Population II stars are older, cooler, less luminous, metal-deficient stars commonly found in globular clusters and the nuclei of galaxies. The initial mass function (IMF) of present-day stars was first parameterized by Ed Salpeter in 1995 as a single power-law for the number of stars $N_\star$ as a function of stellar mass $m_\star$,

$$\frac{dN_\star}{dm_\star} \propto m_\star^{-\alpha}, \tag{12.5}$$

with $\alpha \approx 2.35$ for stars much more massive than the Sun. The related power-law index $\Gamma = \alpha - 1$ relates to the number of stars per $\log m_\star$. Figure 12.8 shows data for the value of Γ as a function of $m_\star$ as well as the derived present-day IMF in different samples of stars.

The dependence of the IMF on the heavy element abundance (metallicity) or redshift is not known. As discussed in §5.2.2, ab-initio simulations of the first metal-free stars, so-called Population III, suggest that they were likely massive with a top-heavy IMF. The corresponding IMF can be simply parameterized by a functional form similar to that of present-day stars but with different characteristic mass and different low-mass and high-mass cut-offs. Since massive stars have short lifetimes, the first massive stars did not survive to the present time and might have been qualitatively distinct from the stars that are observed today. But even after galaxies were enriched with heavy elements, the IMF might have still been redshift dependent because the characteristic density and temperature of the interstellar medium

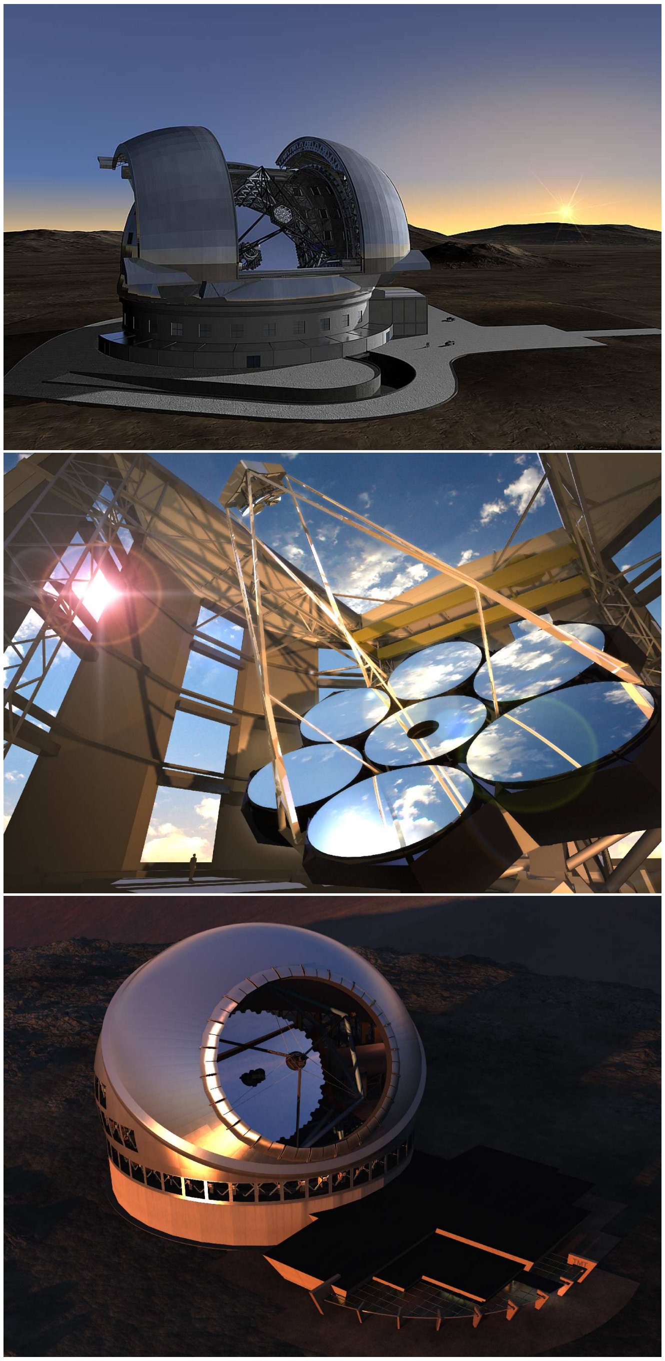

Figure 12.7 Artist's conception of the designs for three future giant telescopes that will be able to probe the first generation of galaxies from the ground: the European Extremely Large Telescope (EELT, top), the Giant Magellan Telescope (GMT, middle), and the Thirty Meter Telescope (TMT, bottom). Images credits: the European Southern Observatory (ESO), the GMT Partnership, and the TMT Observatory Corporation.

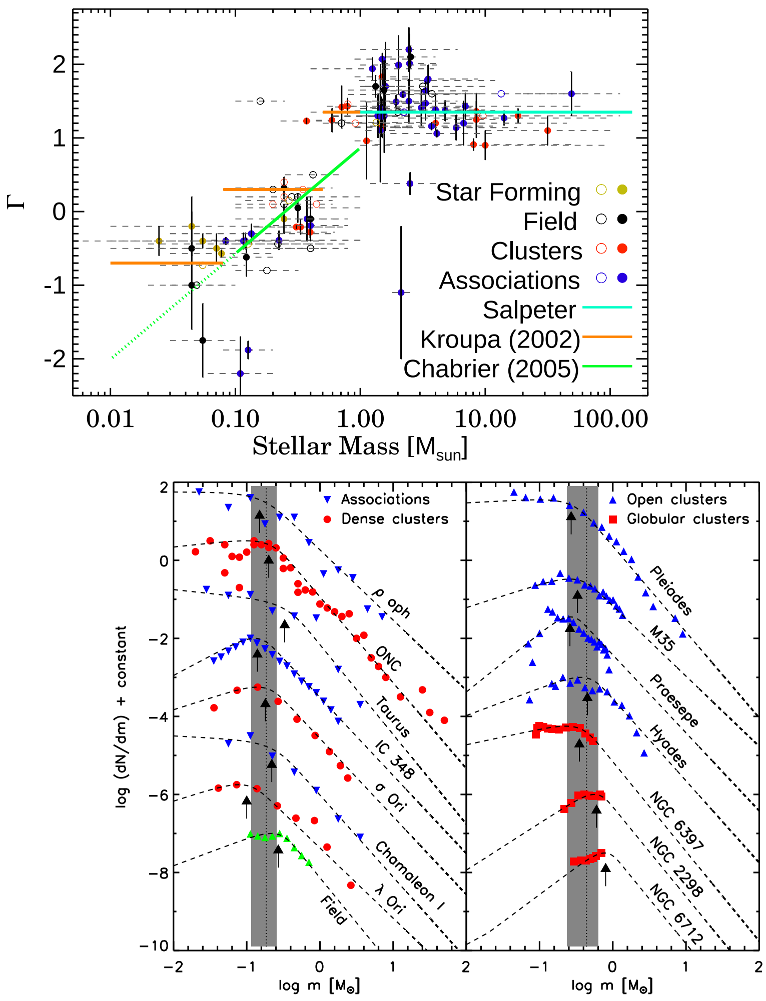

Figure 12.8 *Upper panel:* The derived power-law index, Γ, of the IMF in nearby star forming regions, clusters and associations of stars within the Milky Way galaxy, as a function of sampled stellar mass (points are placed in the center of log m range used to derive each index, with the dashed lines 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).

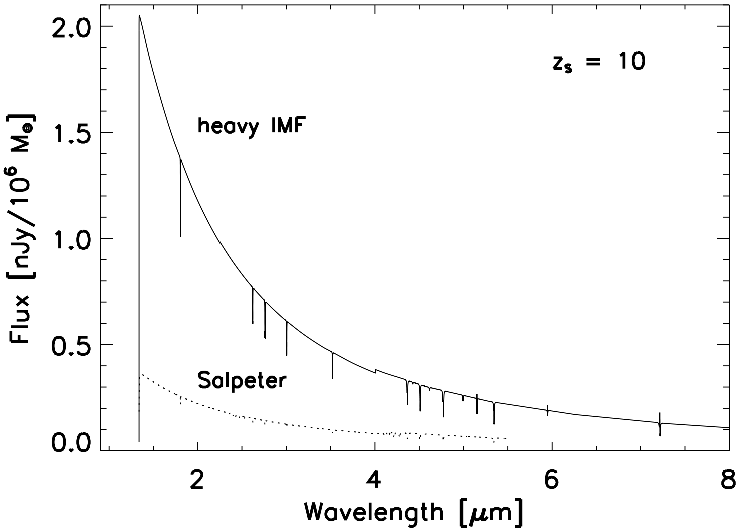

Figure 12.9 Comparison of the observed flux per unit frequency from a galaxy at a redshift $z_s = 10$ for a Salpeter IMF (*dotted line*; Tumlinson, J., & Shull, M. J. *Astrophys. J.* **528**, L65 (2000)) and a purely massive IMF (*solid line*; Bromm, V. Kudritzki, R. P. & Loeb, A. *Astrophys. J.* **552**, 464 (2001)). 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 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 §GPT). For the same total stellar mass, the observable flux is larger by an order of magnitude for stars biased towards having masses $> 100M_\odot$.

were different than they are today. In particular, since the temperature floor of the gas in galaxies is set by the cosmic microwave background at $2.7\ \mathrm{K} \times (1+z)$, it is reasonable to expect that the characteristic mass of stars had been higher at $z \sim 10$ than it is today, irrespective of the metallicity. Spectroscopic observations with JWST will be able to constrain the IMF of the first galaxies, as illustrated in Figure 12.9.

At any given redshift, the brightest among the common stars in galaxies are those stars whose lifetime is comparable to the age of the Universe. While this selects Sun-like stars in the present-day Universe, it favors a stellar mass of $\sim 3M_\odot[(1+z)/10]^{0.6}$ during the epoch of reionization.

12.3 GALAXY EVOLUTION

For a given IMF and metallicity, it is possible to calculate the evolution of the spectral luminosity of a galaxy with time. A library of galaxy spectra computed by G. Bruzual and S. Charlot using their Isochrone Synthesis Spectral Evolutionary Code [137]. A related code, *Starburst99*, for calculating spectra of star-forming galaxies was compiled by Leitherer et al. (1999). [138] The evolution of the luminosity of a galaxy also depends on its star formation rate (SFR) history, $\dot{M}_\star(t)$. The simplest models involve:

- **Constant star formation:** $\dot{M}_\star$ =const, starting at the formation redshift of the galaxy. In this model, the constant SFR equals the stellar mass of the galaxy, $M_\star$, divided by the age of the galaxy, $(t_H - t_F)$. Since this age must be smaller than the age of the Universe, $t_H \approx 10^9\ \mathrm{yr}[(1+z)/7]^{-3/2}$, an observed SFR lower than $M_\star/t_H$ would falsify this model and imply an alternative model in which $\dot{M}_\star$ was higher in the past.

- **Starburst activity:** $\dot{M}_\star = (M_\star/t_\star)\exp\{-(t-t_0)/t_\star\}$ with a short duration $t_\star$ (often much shorter than a Gyr) after the starting time t_0. A startburst may be regarded as instantaneous ($t_\star \rightarrow 0$) if star formation was limited to a period shorter than the evolutionary timescale of the most massive stars (a few million years). Starburst activity is ocassionally triggered by a merger of two galaxies, during which gas is funneled to their centers by tidal torques.

The high-redshift Universe is characterized by frequent mergers whose time average may resemble the steady mode of star formation, but whose instantaneous SFR fluctuates over short time intervals. Although massive galaxies experience roughly one major merger per Hubble time at $z < 2$, low mass galaxies are expected to experience > 4 major mergers per Hubble time at $z > 10$. The excursion set formalism, described in §3.3.1, can be used to quantify the merger rate of galaxies at these early time.

A galaxy luminosity depends on the initial mass function (IMF) of its stars. Most of the UV is emitted by massive stars with a lifetime of 1-10 million years. Therefore the UV luminosity traces the SFR of massive stars, while the infrared emission measures the stellar mass budget of the galaxy.

12.4 METHODS FOR IDENTIFYING HIGH-RESHIFT GALAXIES

Most 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 spectroscopic confirmation. One technique makes use of narrow-band imaging to identify galaxies for which highly-redshifted line emission falls within the selected band. This method is typically applied to the Lyman-α line, whose strength is 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 second observational technique adopts several broad bands to estimate of 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. From the observed spectra of quasars and GRBs at redshifts $z > 6$ it is known that the intergalactic Lyman-α absorption is so high that no flux should be detected just shortward of the observed Lyman-α wavelength 1216Å$(1+z)$ (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 z' (centered at ~ 9000Å) and i' (centered at ~ 8000Å) of HST, as illustrated in Figure 12.10. This method was first used at lower redshifts, $z \sim 3$–4, where the neutral hydrogen column 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).

The key challenge of observers is to obtain a sufficiently high signal-to-noise ratio that the Lyman-α dropout galaxies can be safely identified through the detection of a single redder band. Figure 12.11 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. The Lyman-α line is asymmetric with a sharp cut-off on the blue wing, as expected from intergalactic absorption. The Lyman-α line emission does not emerge from some galaxies. The lengthy trajectory of Lyman-α photons due to resonant scattering makes them particularly vulnerable to absorption by dust, although the level of absorption depends on the geometry and clumpiness of the interstellar medium.[139]

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 instruments 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 HST studies of galaxies at $z \sim 3$ constrained the initial mass function of stars as well as the metallicity and kinematics of the interstellar medium in them.

12.5 LUMINOSITY FUNCTION

The luminosity function (LF) of galaxies, $\phi(L)dL$, provides the number of galaxies per comoving volume within the luminosity bin between L and $L + dL$. A popular

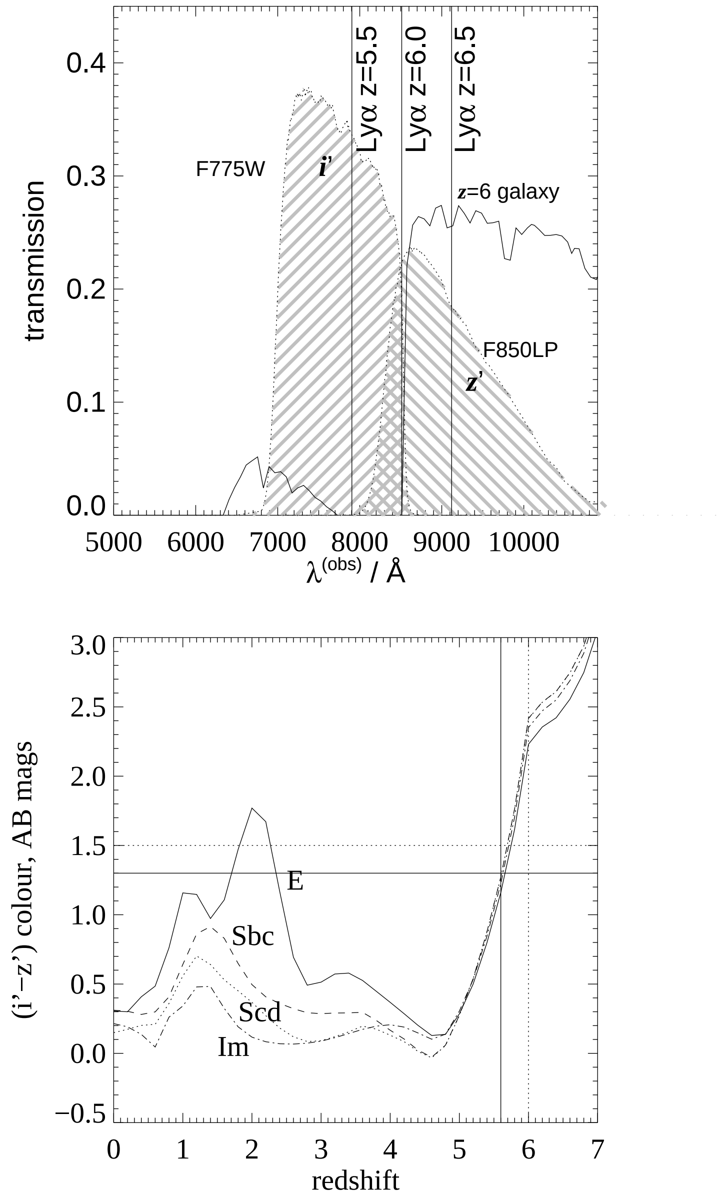

Figure 12.10 *Top panel:* The i' and z' bands of HST (shaded regions) on top of the generic spectrum from a galaxy at a redshift $z = 6$ (solid line). The Lyman-α wavelength at various redshifts is also shown. *Bottom panel:* Models of the color-redshift tracks for different types of galaxies with non-evolving 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).

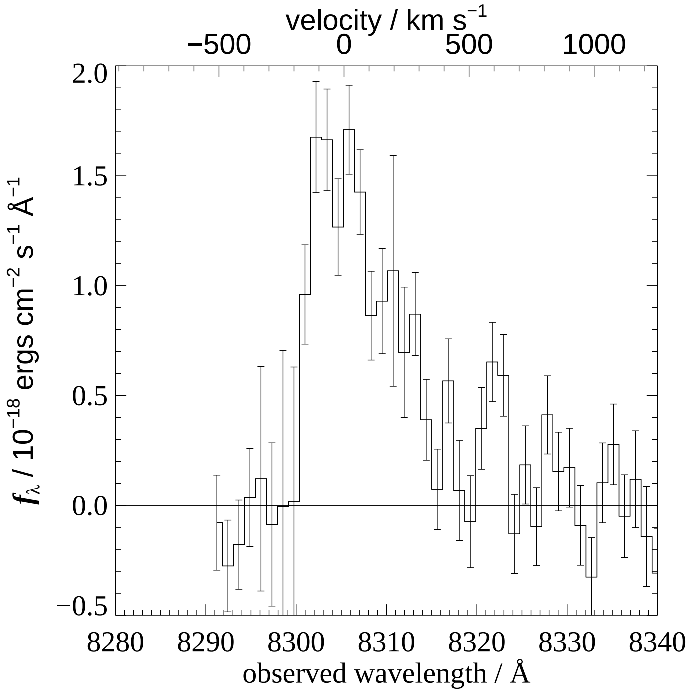

Figure 12.11 The Lyman-α emission line of a typical i'-dropout galaxy SBM03# at $z = 5.83$. Figure credit: Stanway, E., et al. *Astrophys. J.* **607**, 704 (2004).

fitting form is provided by the *Schechter function*,

$$\phi(L) = \phi_\star \left(\frac{L}{L_\star}\right)^{-\alpha} \exp\left(-\frac{L}{L_\star}\right), \tag{12.6}$$

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 ($\sim L_\star$) 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, $u_{\rm gal} = \int_0^\infty \phi(L)LdL = \phi_\star L_\star \Gamma(\alpha+2)$, where Γ is the incomplete 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 because there is a minimum luminosity for galaxies. This minimum is set by the threshold for the assembly of gas into dark matter halos (Jeans mass) or by the cooling threshold of the gas (e.g. $\sim 10^4$ K for atomic hydrogen transitions), required for fragmentation and star formation.

Attempts to reproduce the evolution in the LF 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 LF as a function of redshift) is lacking.

The luminosity function, dn/dL, is related to the comoving density of halos per unit halo mass, dn/dM, discussed in §3.3. A variety of semi-analytic modeling with different levels of complexity can be employed to derive $L(M)$. The simplest model [140] involves two free parameters: *(i)* 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. 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

$$\dot{M}_\star(M) = \frac{f_\star \times (\Omega_b/\Omega_m) \times M}{t_\star}. \tag{12.7}$$

For comparison to data on Lyman-break galaxies (LBGs), the star formation rate can be converted to the luminosity per unit frequency at a rest frame wavelength of 1500 Å based on the relation [141] $L_{1500\mathring{\rm A}} = 8.0\times10^{27}(\dot{M}_\star/M_\odot\ {\rm yr}^{-1})\,{\rm erg\ s^{-1}Hz^{-1}}$. This conversion factor assumes a Salpeter initial mass function (IMF) of stars; if the IMF is more top-heavy than the Salpeter IMF, the far-ultraviolet luminosity will be greater for a given $\dot{M}_\star$. Comparison to surveys of Lyman-α emitters (LAEs) requires a conversion of the star formation rate to a Lyman-α luminosity. This can be achieved by assuming that two-thirds of all recombining photons result in the emission of a Lyman-α photon (so-called, 'case-B' recombination). The ionizing photon production rate is calculated from the star formation rate for a given IMF and metallicity. For a metallicity of 0.05 the solar value and a Salpeter IMF, one finds[142] $N_\gamma = 3 \times 10^{53}$ ionizing photons per second per star formation rate in $M_\odot\ {\rm yr}^{-1}$. A top-heavy Population III IMF produces beyond an order of magnitude more ionizing photons. Since the Lyman-α photons are assumed to be produced via recombinations, only the fraction of ionizing photons which do not escape into the intergalactic medium, $(1 - f_{\rm esc})$, produce Lyman-α photons. Furthermore, only a fraction, $T_{\rm Ly\alpha}$, of the emitted Lyman-α photons escape the galaxy and are transmitted through the intergalactic medium (IGM). With this prescription, the Lyman-α luminosity is related to the halo mass as follows,

$$L_{\rm Ly\alpha} = \frac{2}{3} N_\gamma T_{\rm Ly\alpha} (1 - f_{\rm esc}) \dot{M}_\star. \tag{12.8}$$

A substantial change in the IGM transmission parameter $T_{\rm Ly\alpha}$ is expected to signal the end of reionization.

A more sophisticated model might incorporate the effects of supernova feedback on the luminosity function of star-forming galaxies. Since supernova feedback can significantly reduce the efficiency of star formation in low-luminosity galaxies, it is particularly important to consider when predicting the efficiency of future surveys for high-redshift galaxies aimed at detecting intrinsically fainter systems. Following the scaling relations inferred for local dwarf galaxies,[143] one may define a critical halo mass at each redshift below which the star formation efficiency begins to decrease due to feedback. The star formation efficiency, $\eta(M)$, is a function of halo mass,

$$\eta(M_{halo}) = \begin{cases} f_\star \left(M/M_{\rm crit}\right)^{2/3} & M < M_{\rm crit} \\ f_\star & M > M_{\rm crit} \end{cases} \tag{12.9}$$

where M_{crit} represents the critical halo mass. At a given redshift, the critical halo mass can be related to a critical circular velocity. In the local universe, observations suggest a critical halo velocity of ~ 100 km s^{-1}; the same value may apply at high redshifts if the physics of supernova feedback depends only on the depth of the gravitational potential well of the halos.

12.6 HISTORY OF THE STAR FORMATION RATE DENSITY

The star formation rate (SFR) of a galaxy is commonly gauged based on the following four probes[144] (under the assumption of a 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

$$SFR \approx 1.3 \left(\frac{L_\nu}{10^{28}\ \text{erg s}^{-1}\ \text{Hz}^{-1}}\right)\ M_\odot\ \text{yr}^{-1}. \tag{12.10}$$

- **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,

$$SFR \approx 0.8 \left(\frac{L(\text{H}\alpha)}{10^{41}\ \text{erg s}^{-1}}\right)\ M_\odot\ \text{yr}^{-1}, \tag{12.11}$$

and

$$SFR \approx 1.4 \left(\frac{L(\text{[OII]})}{10^{41}\ \text{erg s}^{-1}} \right)\ M_{\odot}\ \text{yr}^{-1}. \tag{12.12}$$

- **Far-infrared emission** (10–300μm) - measures the total emission from dust heated by young stars,

$$SFR \approx 0.45 \left(\frac{L(\text{FIR})}{10^{43}\ \text{erg s}^{-1}} \right)\ M_{\odot}\ \text{yr}^{-1}. \tag{12.13}$$

- **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,

$$SFR \approx 1.1 \left(\frac{L_{\nu}(1.4\text{GHz})}{10^{28}\ \text{erg s}^{-1}\ \text{Hz}^{-1}} \right)\ M_{\odot}\ \text{yr}^{-1}. \tag{12.14}$$

Intergation 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 (also known as the *Madau Plot*[145]). Figure 12.12 shows the latest determination of this rate based on the UV luminosity function as a function of redshift for all galaxies brighter than $0.07L_{\star}$ at $z = 3$ (corresponding to an AB magnitude of -18.3). In assessing the true history, it is necessary to correct the inferred curve for extinction by dust and feeble galaxies below the detection threshold – which is increasingly important at higher redshifts.

An integral over the luminosity per stellar mass times the cosmic star formation history yields the related radiation background. The fluctuations in this background reflect the clustering and Poisson fluctuations of the associated galaxies.

12.7 GALAXY SEARCHES AND THE IONIZING PHOTON BUDGET DURING REIONIZATION

The production rate of ionizing photons per galaxy, $\dot{N}_{\text{ion}}$, can be derived from the galaxy's luminosity per unit frequency, L_{ν}, above the ionization threshold, $\nu > \nu_{\text{ion}}$ ($= 3.28 \times 10^{15}$ Hz for hydrogen), weighted by the frequency dependence of the cross-section for photo-ionization, $\sigma_{\text{pi}}(\nu)$,

$$\dot{N}_{\text{ion}} = \frac{1}{h} \frac{\int_{\nu_{\text{ion}}}^{\infty} L_{\nu}\sigma_{\text{pi}}(\nu)d\nu}{\int_{\nu_{\text{ion}}}^{\infty} \sigma_{\text{pi}}(\nu)d\nu}. \tag{12.15}$$

Approximating $L_{\nu} \propto \nu^{-\alpha_s}$ and $\sigma_{\text{pi}} \propto \nu^{-3}$, yields $\dot{N}_{\text{ion}} \approx [2/(\alpha_s + 2)]L_{\nu_{\text{ion}}}/h$. The spectral luminosity at an observed frequency L_{obs}, needs to be extrapolated to other frequency $\nu > \nu_{\text{obs}}$ using a template spectrum of the galaxy, in order to derive $\dot{N}_{\text{ion}}$. Note that ionizing photons are produced mainly by massive stars, and so the relation between the global star formation rate and $\dot{N}_{\text{ion}}$ depends on the mass

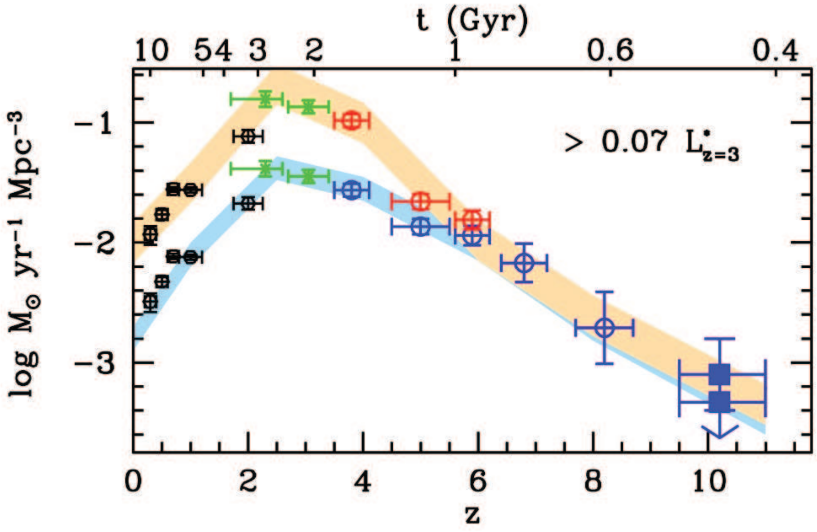

Figure 12.12 The star formation rate density as functions of redshift (lower horizontal axis) and cosmic time (upper axis), for galaxies brighter than -18.3 AB magnitude (corresponding to $0.07 L_\star$ at $z = 3$). The conversion from observed UV luminosity to star formation rate assumed a Salpeter 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/0912.4263v2 (2009).

function of stars. It also depends on the average escape fraction of ionizing photons from galaxies, $f_{\rm esc}$.

If a fraction $f_{\rm esc}(L)$ of the ionizing photons escape from each galaxy of luminosity L, then the average production rate of ionizing photon per comoving volume is given by,

$$\dot{n}_{\rm ion} = \int_0^\infty f_{\rm esc}(L)\dot{N}_{\rm ion}(L)\frac{dn}{dL}dL. \tag{12.16}$$

Before reionization, $\dot{n}_{\rm ion}$ dictates the expansion rate of ionized regions. The Universe can be reionized only when the time integral of this rate gives more than one ionizing photon per proton in the IGM,

$$\int_{z_{\rm reion}}^\infty \dot{n}|\frac{dt}{dz}|dz > 2\times10^{-7}~{\rm cm}^{-3}, \tag{12.17}$$

where $|dt/dz| = [(1+z)H(z)]^{-1} \approx [\sqrt{\Omega_m}H_0]^{-1}(1+z)^{-5/2}$. To produce one ionizing photon per baryon requires a minimum comoving density of Population II stars of,

$$\rho_\star \approx 1.7\times10^6 f_{\rm esc}^{-1}~M_\odot~{\rm Mpc}^{-3}, \tag{12.18}$$

or equivalently, a cosmological density parameter in stars of $\Omega_\star \sim 1.25\times10^{-5}f_{\rm esc}^{-1}$. More typically, the threshold for reionization involves at least a few ionizing photons per proton (with the right-hand-side being $\sim 10^{-6}~{\rm cm}^{-3}$), since the recombination time at the mean density is comparable to the age of the Universe at $z\sim10$.

After reionization, $\dot{n}_{\rm ion}$ balances the average recombination rate per unit comoving volume in the IGM,

$$\dot{n}_{\rm ion} = C\alpha_B\langle n_e(z=0)\rangle^2(1+z)^3, \tag{12.19}$$

where $C = \langle n_e^2\rangle/\langle n_H\rangle^2$ is the volume-averaged clumpiness factor of the electron density up to some threshold overdensity of gas which remains neutral (and is redshift dependent, based on $\dot{N}_{\rm ion}$). The value of $C(z)$ needs to be calculated self-consistently with a numerical simulation that incorporates the empirically-derived function, $\dot{n}_{\rm ion}(z)$. For a Salpeter IMF of Population II stars at solar metallicity, the star formation rate per unit comoving volume that is required for maintaining an ionized IGM according to condition (12.19) is,[146]

$$\dot{\rho_\star} \approx 2\times10^{-3}f_{\rm esc}^{-1}C\left(\frac{1+z}{10}\right)^3~M_\odot~{\rm yr}^{-1}~{\rm Mpc}^{-3}. \tag{12.20}$$

Note that conditions (12.17) and (12.19) place different requirements on the UV luminosity function of galaxies. Some papers interpret incorrectly condition (12.20) as a requirement for achieving reionization, instead of its actual meaning as the requirement for keeping the post-reionization IGM ionized.

The flux sensitivity limit of galaxy searches implies that they are missing the faint end of the luminosity function. At high redshifts, this problem becomes acute. Barkana & Loeb (2000) used the mass function of halos and reasonable assumptions about star formation within them to show that most of the star formation at $z>10$ will be in galaxies an order of magnitude fainter than the few nJy sensitivity limit of JWST.[147] This would limit the utility of galaxy surveys in accounting for the full ionizing photon budget during reionization. Alternative probes, such as 21-cm tomography, could calibrate the ionizing photon budget indirectly, based on the growth rate of ionized regions during reionization.

12.8 BIASED CLUSTERING OF HIGH-REDSHIFT GALAXIES

Since high-redshift galaxies represent rare high peaks in the underlying density field, their clustering is expected to be enhanced or 'biased'. The bias parameter, defined as the enhancement in the fluctuation amplitude of the number density of galaxies relative to the matter density, could be a function of halo mass (in liner theory) and spatial scale (when non-linear effects are accounted for). The dependence of clustering on halo mass can be used to infer halo masses statistically from galaxy surveys.

Most simply, the bias can be calculated from linear theory where it is dictated by the underlying clustering of peaks in the density field. The likelihood of observing a galaxy at a random location is proportional to the local number density of its corresponding halos. Given a large scale overdensity δ of comoving radius R, the observed overdensity of galaxies is written as

$$\delta_{\rm g} = b_{\rm g}(M,z)\delta. \tag{12.21}$$

In the Press-Schechter formalism, the bias factor $b_{\rm g}$ is simply the ratio between the number density of halos in a large-scale region of overdensity δ and the number density of halos in the background universe.[148] For small values of $|\delta| << 1$,

$$\begin{aligned} 1+\delta b_{\rm g} &= (1+\delta)\left[\frac{dn}{dM}(\bar{\nu}) + \frac{d^2n}{dMd\nu}(\bar{\nu})\frac{d\nu}{d\delta}\delta\right]\left[\frac{dn}{dM}(\bar{\nu})\right]^{-1} \\ &\approx 1+\delta\left[1+\frac{(\bar{\nu}^2-1)}{\sigma(R)\bar{\nu}}\right], \end{aligned} \tag{12.22}$$

where $\nu = (\delta_{\rm crit}-\delta)/[\sigma(M)]$, $\bar{\nu} \equiv \nu_c = \delta_{\rm crit}/[\sigma(M)]$, $\delta_{\rm crit}$ is the critical linear overdensity for collapse, and $\sigma(M)$ is the variance of the density field smoothed with a top-hat window of radius R corresponding to the halo mass M (where R is much smaller than the size of the region associated with the large-scale overdensity). Here, $(dn/dM)(\bar{\nu})$ and $(dn/dM)(\nu)$ are the average and perturbed mass functions of halos. The Press-Schechter bias is therefore,

$$b_{\rm g}^{\rm PS}(M,z) = 1+\frac{1}{\delta_{\rm crit}}\left[\bar{\nu}^2-1\right]. \tag{12.23}$$

The value of bias $b_{\rm g}$ for a halo mass M may be calculated more accurately based on the Sheth-Tormen extension of the Press-Schechter formalism that accounts for non-spherical collapse,[149]

$$\begin{aligned} b_{\rm g}^{\rm ST}(M,z) = 1+\frac{1}{\delta_{\rm crit}}\Big[&\nu'^2 + b\nu'^{2(1-c)} \\ &- \frac{\nu'^{2c}/\sqrt{a}}{\nu'^{2c}+b(1-c)(1-c/2)}\Big], \end{aligned} \tag{12.24}$$

where $\nu' \equiv \sqrt{a}\bar{\nu}$, $a = 0.707$, $b = 0.5$ and $c = 0.6$. Within linear theory, the bias in equations (12.22) and (12.24) is a function of halo mass, but not of overdensity or spatial scale.

The clustering of galaxies on scales of up to ~ 100 comoving Mpc affects the evolution and size distribution of HII regions during reionization. Since ionized regions form around groups of galaxies, the distribution of neutral hydrogen is expected to be anti-correlated with the distribution of galaxies.[150] Moreover, because of Lyman-α scattering, the gravitational clustering of Lyman-α emitting galaxies is expected to be further modulated by the punctuated distribution of hydrogen in the IGM.[151] The strong bias of early galaxies implies that the metal enrichment of the IGM is highly inhomogeneous, with pockets of prestine (metal-poor) gas left over at late times, capable of forming post-reionization galaxies with Population-III stars.

Clustering also affects the statistical uncertainty in number counts of galaxies within surveys of limited volume, the so-called *cosmic variance*. The total variance in a survey is the sum of contributions from cosmic variance and Poisson shot noise,

$$\frac{\sigma^2_{\rm tot}}{\langle N\rangle^2} = \frac{\langle N^2\rangle - \langle N\rangle^2}{\langle N\rangle^2} = \frac{\sigma^2_{\rm hh}}{\langle N\rangle^2} + \frac{1}{\langle N\rangle}, \tag{12.25}$$

where $\langle N\rangle$ is the mean number of galaxies within the survey volume. The cosmic (or sample) variance, which is the first term on the right hand side of equation (12.25), results from the survey field sometimes lying in a region of high galaxy density and sometime lying in an under-dense region or a void. This contribution can be calculated from linear perturbation theory (i.e. based on the linear power-spectrum, $P(k)$, evaluated at $z = 0$ with $D(z = 0) = 1$) as a function of the minimum halo mass, M, hosting observed galaxies as,

$$\sigma^2_{hh}(M,z) = \frac{(b(>M)\,D(z))^2}{(2\,\pi)^3}\int P(k)\,W^2_{xyz}\,d^3k, \tag{12.26}$$

where $W_{xyz} = W(k_x)\,W(k_y)\,W(k_w)$, $k = \sqrt{k_x^2 + k_y^2 + k_w^2}$, $D(z)$ is the linear growth factor evaluated, for simplicity, at the midpoint of the given redshift range, and $b(>M)$ is the linear bias integrated over all masses above M and weighted by the halo mass function. The window function, $W(k_i)$, is the Fourier transform of a top-hat in the i-th dimension and is given by:

$$W(k_i) = \frac{\sin(k_i\,a_i/2)}{k_i\,a_i/2}. \tag{12.27}$$

In equation (12.27), $a_x = a_y = (\theta/1')\,(\pi/180 \times 60')\,\chi(z)$ are the narrow dimensions of the skewer-shaped survey of angular width θ evaluated at z. The subscript w refers to coordinates along the line-of-sight. The length of the survey, a_w, is given by the comoving distance between the limits bracketting the redshift range of interest. 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.

Figure 12.13 compares the contributions from cosmic and Poisson variance as calculated by linear theory (equation 12.25 and 12.26) for $a_x = a_y = 3.4'$ as a function of the opening angle of the survey, $\theta = a_x/\chi(z)$. This plot can be used to calculate the effectiveness of future surveys with large fields of view. Figure 12.14 shows the variance for surveys of Lyman-break galaxies (LBGs) in the redshift range $z = 6 - 8$. In both figures, adopted from Munoz et al. (2009), each

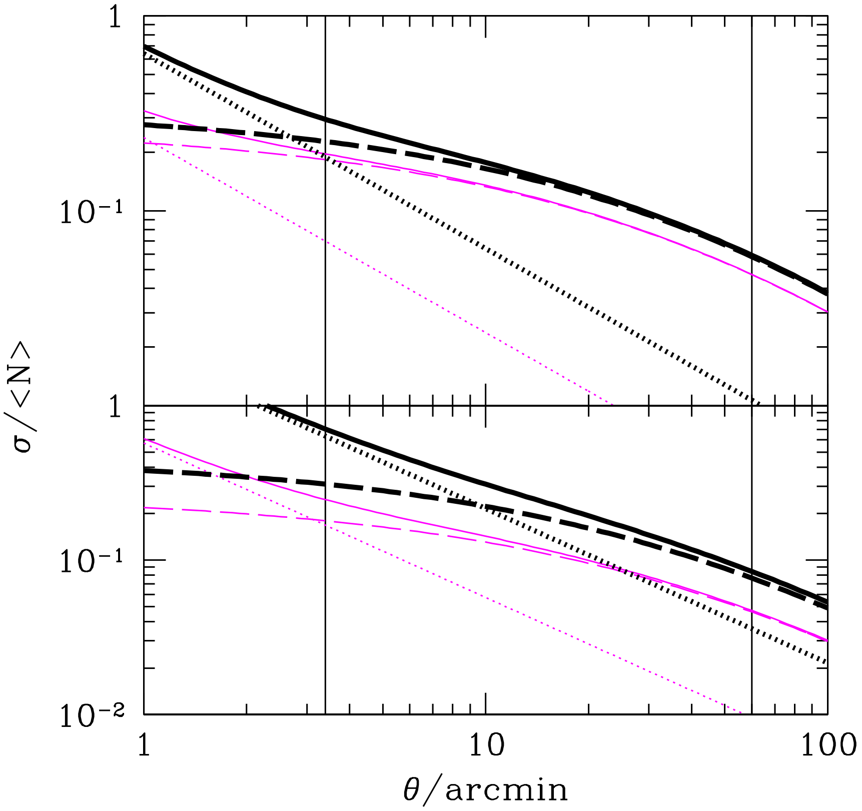

Figure 12.13 The theoretically predicted contributions to the total variance (equation 12.25; solid lines) in LBG dropout surveys as a sum of cosmic variance (dashed lines) and Poisson shot noise (dotted lines) contributions (i.e. the first and second terms, respectively, on the right-hand-side of equation 12.25). 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.*, in press (2010).

halo is assigned an LBG luminosity based on a simple prescription with a duty cycle of 25% and star formation efficiency of 16%. When compared to numerical simulations, the galaxy count statistics are well approximated by the linear-theory expressions at the low luminosity limits.

One may define the skewness in galaxy statistics as the third moment of the probability distribution normalized by the variance to the $3/2$ power:

$$s_3 = \frac{\left\langle (N - \langle N \rangle)^3 \right\rangle}{(\sigma^2)^{3/2}}. \tag{12.28}$$

The skewness as a function of minimum luminosity is presented in the bottom panel of figure 12.14. The seemingly large amplitude variations in the skewness at low luminosity for $z = 6$–8 are due to small numerical fluctuations around the nearly zero skewness from numerical simulations, plotted on a log scale. The numerical simulations indicate that the probability distribution of bright LBGs 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 nonlinear clustering on the small scales probed by the narrowness of the survey skewer.

12.9 MOLECULES, DUST AND THE INTERSTELLAR MEDIUM

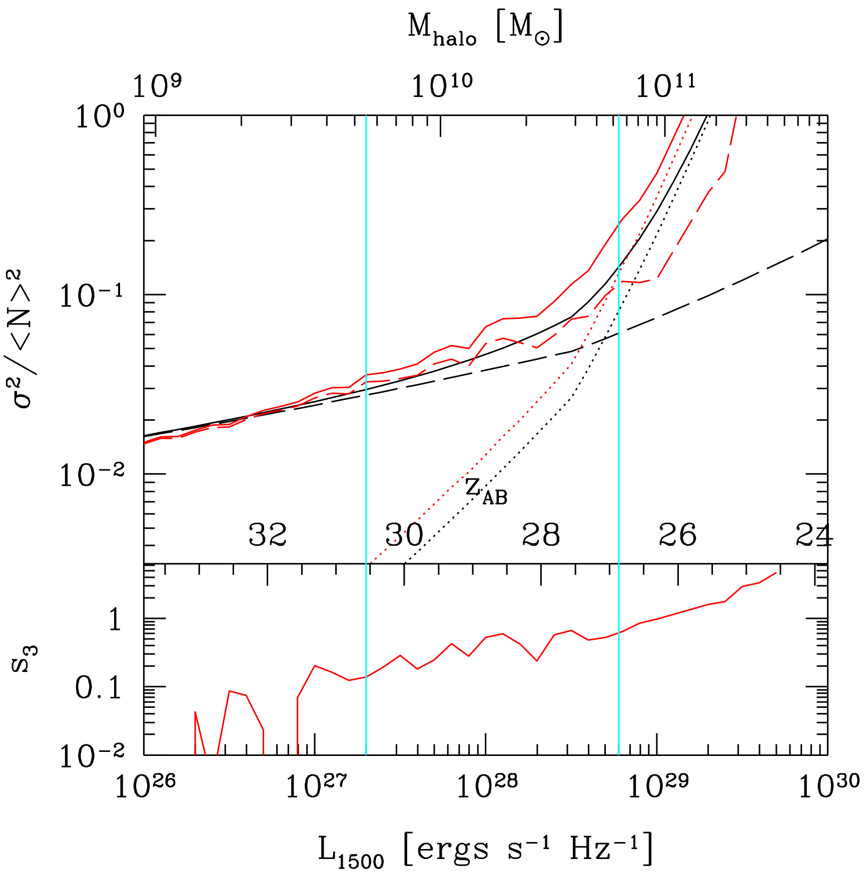

Figure 12.14 **Upper panel:** predicted relative contributions to the fractional variance in the number counts of galaxies as a function of UV luminosity at an emission wavelength of 1500Å, z-band AB magnitude, or host halo mass in counts of LBGs within a dropout survey spanning the redshift interval $z = 6$–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 sample variance and Poisson noise, respectively. Upper curves show the results from numerical simulations, while lower curves were calculated analytically based on linear perturbation theory. Vertical lines brackett the region where the variance is higher than expected due to the skewness of the full count probability distribution but is not Poisson-dominated. **Lower panel:** the skewness of the full galaxy count probability distribution calculated from a numerical simulation based on equation (12.28). Figure credit: Munoz, J., Trac, H., & Loeb, A. *Mon. Not. R. Astron. Soc.*, in press (2010).

Chapter Thirteen

Other Observational Probes of the First Galaxies

13.1 THE COSMIC MICROWAVE BACKGROUND

13.2 THE INFRARED BACKGROUND

13.3 LOW-REDSHIFT SIGNATURES

13.3.1 Clues from the IGM (Thermal History, Metal Enrichment)

13.3.2 The Fossil Record of the Local Group

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] For advanced reading, see Mukhanov, V. *Physical Foundations of Cosmology*, Cambridge University Press, Cambridge (2005).

[4] http://public.web.cern.ch/public/en/LHC/LHC-en.html

[5] Loeb, A., Ferrara, A., & Ellis, R. S. *First Light in the Universe*, Saas-Fee Advanced Course **36**, Springer, New-York (2008), and references therein.

[6] Peebles, P. J. E. *Principles of Physical Cosmology*, Princeton University Press, (1993), p. 626.

[7] 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.

[8] Haiman, Z., Thoul, A. A., & Loeb, A. *Astrophys. J.* **464**, 523 (1996).

[9] Barkana, R., & Loeb, A. *Astrophys. J.* **523**, 54 (1999).

[10] Gnedin, N. Y., & Hui, L. *Mon. Not. R. Astron. Soc.* **296**, 44 (1998).

[11] Loeb, A., & Zaldarriaga, M. *Phys. Rev.* **D71**, 103520 (2005).

[12] Barkana, R., & Loeb, A. *Astrophys. J.* **531**, 613 (2000), and references therein.

[13] Press, W. H., & Schechter, P. *Astrophys. J.* **187**, 425 (1974).

[14] Bond, J. R., Cole, S., Efstathiou, G, & Kaiser, N. *Astrophys. J.* **379**, 440 (1991).

[15] Lacey, C. G., & Cole, S. *Mon. Not. R. Astr. Soc.* **262**, 627 (1993).

[16] Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **476**, 458 (1997).

[17] Hirata, C. M., & Padmanabhan, N. *Mon. Not. R. Astron. Soc.* **372**, 1175 (2006).

[18] Stecher, T. P., & Williams, D. A. *Astrophys. J.* **149**, L29 (1967).

[19] Bromm, V., & Larson, R. B. *Ann. Rev. Astron. & Astrophys.* **42**, 79 (2004), and references therein.

[20] Wyithe, J. S. B., & Loeb, A. *Nature* **441**, 322 (2006).

[21] Muñoz, J. A., Madau, P, Loeb, A., & Diemand, J. *Mon. Not. R. Astron. Soc.*, in press (2009), and references therein.

[22] Turk, M. J., Abel, T., & O'Shea, B *Science*, in press (2009), http://arxiv.org/abs/0907.2919; Stacey, A., Greif, T. H, & Bromm, V. *Mon. Not. R., Astr. Soc.* (2009), http://arxiv.org/abs/0908.0712, and references therein.

[23] Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **391**, 457 (2008).

[24] Bromm, V., & Loeb, A. *New Astron.* **9**, 353 (2004).

[25] Pudritz, R. E. *Science* **295**, 68 (2002), and references therein.

[26] Salpeter, E. *Astrophys. J.* **121**, 161 (1955).

[27] Rees, M. J. *Mon. Not. R. Astron. Soc.* **176**, 483 (1976).

[28] Bromm, V., & Loeb, A. *Nature* **425**, 812 (2003).

[29] Furlanetto, S. R., & Loeb, A. *Astrophys. J.* **556**, 619 (2001).

[30] Furlanetto, S. R., & Loeb, A. *Astrophys. J.* **588**, 18 (2003).

[31] Miralda-Escudé, J., & Rees, M. *Astrophys. J.* **478**, L57 (2007).

[32] Mackey, J., Bromm, V. & Hernquist, L. *Astrophys. J.* **586**, 1 (2003); Johnson, J., & Bromm, V. *Mon. Not. R. Astron. Soc.* **366**, 247 (2006); McKee, C. F. & Tan, J. *Astrophys. J.* **681**, 771 (2008); Greif, T. H. et al. *Mon. Not. R. Astron. Soc.* **387**, 1021 (2008).

[33] Wood, K., & Loeb, A. *Astrophys. J.* **545**, 86 (2000).

[34] Fukugita, M., Hogan, C. J., & Peebles, P. J. E. *Astrophys. J.* **503**, 518 (1998).

[35] See, e.g. Barkana, R., & Loeb, A. *Astrophys. J.* **539**, 20 (2000); Stark, D. P., Loeb, A., & Ellis, R. S. *Astrophys. J.* **668**, 627 (2007), and references therein.

[36] Mo, H. J., & White, S. D. M. *Mon. Not. R. Astron. Soc.* **336**, 112 (2002).

[37] Heger, A., & Woosley, S. E. *Astrophys. J.* **567**, 532 (2002); Heger, A., et al. *Astrophys. J.* **591**, 288 (2003).

[38] Frebel, A., Johnson, J. L., & Bromm, V. *Astrophys. J.* **392**, L50 (2009), and references therein.

[39] Mo H. J., Mao, S., & White, S. D. M. *Mon. Not. R. Astron. Soc.* **295**, 319 (1998).

[40] Warren, M. S., Quinn, P. J., Salmon, J. K., & Zurek, W. H. *Astrophys. J.* **399**, 405 (1992).

[41] Schmidt, M. *Astrophys. J.* **129**, 243 (1959); **137**, 758 (1963); Kennicutt, R. C. *Astrophys. J* **498**, 541 (1998); *Proc. Astron. Soc. Pac.* **390**, 149 (2008).

[42] For recent reviews about GRBs, see Mészáros, P. *AIP Conf. Proc.* **924**, 3 (2007) & *Rep. Prog. Phys.* **69**, 2259 (2006) ; Piran, T. *Nuovo Cimento* **B 121**, 1039 (2006) & *Rev. Mod. Phys.* **76**, 1143 (2005); and Gehrels, N., Ramirez-Ruiz, E., & Fox, D. B. *Ann. Rev. Astron. & Astrophys.* **47**, 567 (2009).

[43] Zhang, W., Woosley, S. E., & MacFadyen, A. I. *J. of Phys. Conf. Ser.* **46**, 403 (2006)

[44] http://swift.gsfc.nasa.gov/

[45] Tanvir, N. R., et al. *Nature* **461**, 1254 (2009); Salvaterra, R., et al. *Nature* **461**, 1258 (2009).

[46] Bardeen et al. (1972)

[47] Shakura, N. I., & Sunyaev, R. A. *Astron. & Astrophys.* **24**, 337 (1973); Novikov, I. D., & Thorne, K. S. in *Black Holes*, C. DeWitt and B. DeWitt, editors, Gordon and Breach, New York, New York (1973).

[48] Velikhov, E. P. *J. Exp. Theor. Phys.* **36**, 1398 (1959); Chandrasekhar, S. *Proc. Nat. Acad. Sci/* **46**, 253 (1960); Acheson, D. J., & Hide, R. *Rep. Prog. Phys.* **36**, 159 (1973); Balbus, S. A., & Hawley, J. F. *Astrophys. J.* **376**, 214 (1991).

[49] Haiman, Z., Kocsis, B., & Menou, K. *Astrophys. J.* **700**, 1952 (2009).

[50] Binney, J., & Tremaine, S., *Galactic Dynamics* (2nd edition), Princeton University Press (2008), p. 362.

[51] Narayan, R. & McClintock, J. *New Astronomy Reviews* **51**, 733 (2008), and refs therein.

[52] Begelman, M. C., Blandford, R. D., & Rees, M. J. *Rev. Mod. Phys.* **56**, 255 (1984).

[53] Wyithe, J. S. B., & Loeb, A. *Astrophys. J.* **595**, 614 (2003).

[54] A. Laor & B. Draine *Astrophys. J.* **402**, 441, (1993).

[55] H. Netzer & A. Laor *Astrophys. J.* **404**, L51, (1993).

[56] See Loeb, A., http://arxiv.org/abs/0909.0261 (2009), and references therein.

[57] Pretorius, F. *Phys. Rev. Lett.* **95**, 121101 (2005); Campanelli, M. et al. *Phys. Rev. Lett.* **96**, 111101 (2006); Baker, J. et al. *Phys. Rev. Lett.* **96**, 111102 (2006).

[58] Loeb, A. *Phys. Rev. Lett.* **99**, 041103 (2007).

[59] Blecha, L., & Loeb, A. *Mon. Not. R. Astr. Soc.* **390**, 1311 (2008); Tanaka, T., & Haiman, Z. *Astrophys. J.* **696**, 1798 (2009).

[60] Gebhardt, K., et al. *Astrophys. J.* **539**, L13 (2000); Ferrarese, L., & Merritt, D. *Astrophys. J.* **539**, L9 (2000). These two competing papers appeared simultaneously since I suggested the topic to their lead authors at the same time.

[61] Magorrian, J., et al. *Astron. J.* **115**, 2285 (1998).

[62] Silk, J., & Rees, M. J. *Astron. & Astrophys.* **331**, L1 (1998).

[63] Milosavljević, M., & Loeb, A. *Astrophys. J.* **604**, L45 (2004).

[64] Dietrich, M., & Hamann, F. *Rev. Mex. de Astron. Astrof. Conf. Ser.* **32**, 65 (2008).

[65] See, e.g. Wyithe, J. S. B., & Loeb, A. *Astrophys. J.* **595**, 614 (2003); Hopkins, P. F., & Hernquist, L. *Astrophys. J.* **698**, 1550 (2009).

[66] Genzel, R., & Karas, V. IAU Symp. **238**, 173 (2007); Ghez, A., et al. *Astrophys. J.* **689**, 1044 (2008).

[67] M. Colpi, L. Mayer, & F. Governato *Astrophys. J.* **525**, 720 (1999).

[68] Hoffman, L., & Loeb, A. *Mon. Not. R. Astron. Soc.* **377**, 957 (2007).

[69] Dunkley, J., et al. *Astrophys. J. Suppl.* **180**, 306 (2009).

[70] Faucher-Giguère, C.-A., Lidz, A., Hernquist, L., & Zaldarriaga, M. *Astrophys. J.* **682**, L9 (2008); Wyithe, J. S. B., & Loeb, A. *Astrophys. J.* **586**, 693 (2003).

[71] Strömgren, B. *Astrophys. J.* **89**, 526 (1939).

[72] Shapiro, P. R., & Giroux, M. L. *Astrophys. J.* **321** L107 (1987).

[73] Barkana, R., & Loeb, A. *Phys. Rep.* **349**, 129 (2001), and references therein.

[74] Wyithe, J. S. B., & Loeb, A. *Nature* **427**, 815 (2004); *Astrophys. J.* **610**, 117 (2004).

[75] Gnedin, N. *Astrophys. J.* **535**, 530 (2000).

[76] Barkana, R., & Loeb, A., *Astrophys. J.* **609**, 474 (2004).

[77] Furlanetto, S. R, Zaldarriaga, M., & Hernquist, L. *Astrophys. J.* **613**, 1 (2004).

[78] Zahn, O., et al. *Astrophys. J.* **654**, 12 (2007).

[79] Kaiser, N. *Astrophys. J.* **284**, L9 (1984).

[80] Wyithe, J. S. B., & Loeb, A. *Nature* **432**, 194 (2004).

[81] Barkana, R., & Loeb., A. *Astrophys. J.* **539**, 20 (2000).

[82] Babich, D., & Loeb, A. *Astrphys. J.* **640**, 1 (2006); Wyithe, S.J., & Loeb, A. *Mon. Not. R. Astron. Soc* **382**, 921 (2007).

[83] Furlanetto, S.R., & Loeb, A. *Astrophys. J.* **588**, 18 (2003); **634**, 1 (2005).

[84] Scheuer, P. A. G. *Nature* **207**, 963 (1965).

[85] Gunn, J. E., & Peterson, B. A. *Astrophys. J.* **142**, 1633 (1965).

[86] Miralda-Escudé, J. *Astrophys. J.* **501**, 15 (1998).

[87] Verner, D. A., Ferland, G. J., Korista, T., & Yakovlev, D. G. *Astrophys. J.* **465**, 487 (1996).

[88] Loeb, A., & Rybicki, G. *Astrophys. J.* **524**, 527 (1999) and **520**, L79 (1999); see also Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **386**, 492 (2008).

[89] Steidel, C. C., et al. *Astrophys. J.* **532**, 170 (2000).

[90] Matsuda, Y., et al. *Astron. J.*, **128**, 569 (2004); Matsuda, Y., et al. *Astrophys. J.*, **640**, L123 (2006); Saito, T., et al. *Astrophys. J.* **537**, L5 (2000); Yang, Y., et al. *Astrophys. J.* **693**, 1579 (2009).

[91] Chapman, S. C., et al. *Astrophys. J.*, **548**, L17 (2001); Chapman, S. C., et al. *Mon. Not. R. Astron. Soc.*, **363**, 1398 (2005); Geach, J. E., et al. *Astrophys. J.*, **640**, L123 (2006).

[92] Geach, J. E., et al. *Astrophys. J.*, **655**, L9 (2007); Basu-Zych, A., & Scharf, C., *Astrophys. J.*, **615**, L85 (2004).

[93] Nilsson, K. K., et al. *A&A.*, **452**, L23 (2006); Smith, D. J. B., et al. *Mon. Not. R. Astron. Soc.*, **378**, L49 (2007).

[94] Ouchi, M. et al. *Astrophys. J.* **696**, 1164O (2009).

[95] Haiman, Z., Spaans, M., & Quataert, E. *Astrophys. J.* **537**, L5 (2000); Fardal, M. A., et al. *Astrophys. J.* **562**, 605 (2001).

[96] Haiman, Z., & Rees, M. J. *Astrophys. J.*, **556**, 87 (2001).

[97] Scharf, C. et al. *Astrophys. J.*, **596**, 105 (2003).

[98] Mori., M., et al. *Astrophys. J.*, **613**, L97 (2004).

[99] Rees, M. J. *Mon. Not. R. Astron. Soc.*, **239**, 1P (1989).

[100] Dijkstra, M. & Loeb, A. *Mon. Not. R. Astron. Soc.* **400**, 1109 (2009); Goerdt, T., et al., preprint arXiv 0911.5566 (2009); Faucher-Giguére, C.-A., et al. preprint arXiv:1005.3041 (2010).

[101] Keres, D., et al. *Mon. Not. R. Astron. Soc.*, **363**, 2 (2005).

[102] Keres, D., et al., *Mon. Not. R. Astron. Soc.***395**, 160 (2009).

[103] For further reading on 21-cm cosmology, see Furlanetto, S. R., Oh, S. P., & Briggs, F. H. *Phys. Rep.* **433**, 181 (2006), and references therein.

[104] Loeb, A., & Zaldarriaga, M. *Phys. Rev. Lett.* **92**, 211301 (2004).

[105] Wouthuysen, S. A. *Astron. J.* **57**, 31 (1952); Field, G. B. *Proc. IRE* **46**, 240 (1958).

[106] Madau, P., Meiksin, A., & Rees, M. J. *Astrophys. J.* **475**, 429 (1997).

[107] Scott, D., & Rees, M. J. *Mon. Not. R. Astron. Soc.* **247**, 510 (1990).

[108] http://www.lofar.org/

[109] http://www.haystack.mit.edu/ast/arrays/mwa/site/index.html

[110] http://arxiv.org/abs/astro-ph/0502029

[111] http://arxiv.org/abs/0904.2334

[112] *http://www.skatelescope.org*

[113] See, e.g. Bowman, J. D., Rogers, A. E. E., & Hewitt, J. N. *Astrophys. J.* **676**, 1 (2008).

[114] See, e.g. Lidz, A., Zahn, O., McQuinn, M., Zaldarriaga, M., & Hernquist, L. *Astrophys. J.* **680**, 962 (2008), and references therein.

[115] Wyithe, J. S. B. & Loeb, A. *Mon. Not. R. Astr. Soc.* **383**, 1195 (2008).

[116] Barkana, R., & Loeb, A. *Astrophys. J.* **624**, L65 (2005).

[117] Loeb, A., & Wyithe, J. S. B. *Phys. Rev. Lett.* **100**, 161301 (2008); Mao, Y., et al. *Phys. Rev.* **D78**, 023529 (2008).

[118] 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).

[119] For an overview of the current observational status, see Ellis, R. S. (2007), http://arxiv.org/abs/astro-ph/0701024.

[120] Thompson, R. I. *Astrophys. J.* **596**, 748 (2003).

[121] Bouwens, R. J., Illingworth, G. D., Franx, M., & Ford, H. *Astrophys. J* **686**, 230 (2008).

[122] See, e.g. Kashikawa, N., et al. *Astrophys. J.* **648**, 7 (2006); Dawson, S., et al. *Astrophys. J.* **671**, 1227 (2007).

[123] See, e.g. Yan, H., et al. *Astrophys. J.* **651**, 24 (2006); Rhoads, J. E., et al. *Astrophys. J.* **697**, 942 (2009); Ouchi, M., et al., http://arxiv.org/abs/0908.3191 (2009).

[124] Ota, K., et al. *Astrophys. J.* **677**, 12 (2008); Stark, D. et al. http://arxiv.org/abs/1003.5244 (2010).

[125] Bradley, L., et al. *Astrophys. J.* **678**, 647 (2008).

[126] Stark, D., et al. *Astrophys. J.* **663**, 10 (2007); Bouwens, R. J., et al. *Astrophys. J* **690**, 1764 (2009).

[127] Ciardi, B., & Loeb, A. *Astrophys. J.* **540**, 687 (2000).

[128] Barkana, R., & Loeb, A. *Astrophys. J.* **601**, 64 (2004).

[129] Bromm, V., & Loeb, A. *Astrophys. J.* **642** 382 (2006).

[130] See, e.g. Barkana, R., & Loeb, A. *Astrophys. J.* **531**, 613 (2000) and **539**, 20 (2000).

[131] Bromm, V., Kudritzki, R. P., & Loeb, A. *Astrophys. J.* **552**, 464 (2001).

[132] http://www.eso.org/sci/facilities/eelt/

[133] http://www.gmto.org/

[134] http://www.tmt.org/

[135] Wyithe, J. S. B., & Loeb, A. *Mon. Not. R. Astron. Soc.* **375**, 1034 (2007).

[136] http://almaobservatory.org/

[137] See **http://www.cida.ve/∼bruzual/bcXXI.html** and **http://www2.iap.fr/users/charlot/bc2003/**.

[138] Leitherer, C., et al. *Astrophys. J. Suppl.* **123**, 3 (1999); see http://www.stsci.edu/science/starburst99/

[139] Neufeld, D. *Astrophys. J.* **370**, L85 (1991).

[140] Wyithe, J. S. B., & Loeb, A. *Nature* **441**, 322 (2006); Stark, D. P., Loeb, A., & Ellis, R. S. *Astrophys. J.* **668**, 627 (2007).

[141] Madau, P., Pozzetti, L., & Dickinson, M. *Astrophys. J.* **498**, 106 (1998).

[142] Schaerer, D. *Astr. & Astrophys.* **397**, 527 (2003).

[143] Dekel, A., & Woo, J. *Mon. Not. R Astr. Soc.* **344**, 1131 (2003).

[144] Ellis, R. S. in *First Light in the Universe*, SAAS-Fee Advanced Course **36**, Springer, New-York (2008).

[145] Madau, P. et al. *Mon Not. R. Astron. Soc.* **283**, 1388 (1996).

[146] Madau, P., Haardt, F., & Rees, M. J., *Astrophys J.* **514**, 648 (1999).

[147] Barkana, R., & Loeb, A. *Astrophys. J.* **539**, 20 (2000); see also, Salvaterra, R., Ferrara, A., & Dayal, P. *Mon. Not. R. Astron. Soc.*, submitted, arXiv:1003.3873 (2010).

[148] Mo, H. J., & White, S. D. M. *Mon. Not. R. Astron. Soc.* **282**, 347 (1996).

[149] Sheth, R., Mo, H. J., & Tormen, G. *Mon. Not. R. Asytron. Soc.* **323**, 1 (2001).

[150] Wyithe, J. S. B., Loeb, A., & Schmidt, B. P. *Mon. Not. R. Astron. Soc.* **380**, 1087 (2007); Furlanetto, S. R. & Lidz, A. *Astrophys. J.* **660**, 1030 (2007).

[151] McQuinn, M. Hernquist, L., Zaldarriaga, M., & Dutta, S. *Mon. Not. R. Astron. Soc.* **381**, 1101 (2008).

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)	=	6.67×10^{-8} cm^3 g^{-1} s^{-2}
Speed of light (c)	=	3.00×10^{10} cm s^{-1}
Planck's constant (h)	=	6.63×10^{-27} erg s
Electron mass (m_e)	=	9.11×10^{-28} g $\equiv 511\,\mathrm{keV}/c^2$
Electron charge (e)	=	4.80×10^{-10}esu
Proton mass (m_p)	=	1.67×10^{-24} g $= 938.3\,\mathrm{MeV}/c^2$
Boltzmann's constant (k_B)	=	1.38×10^{-16} erg K^{-1}
Stefan-Boltzmann constant (σ)	=	5.67×10^{-5} erg cm^{-2} s^{-1} K^{-4}
Radiation constant (a)	=	7.56×10^{-15} erg cm^{-3} K^{-4}
Thomson cross-section (σ_T)	=	6.65×10^{-25} cm^2

Astrophysical numbers		
Solar mass ($M_\odot$)	=	1.99×10^{33} g
Solar radius ($R_\odot$)	=	6.96×10^{10} cm
Solar luminosity ($L_\odot$)	=	3.9×10^{33} erg s^{-1}
Hubble constant today (H_0)	=	$100h$ km s^{-1} Mpc^{-1}
Hubble time (H_0^{-1})	=	$3.09 \times 10^{17} h^{-1}$ s $= 9.77 \times 10^9 h^{-1}$ yr $\equiv 3h^{-1}$ Gpc$/c$
critical density (ρ_c)	=	$1.88 \times 10^{-29} h^2$ g cm^{-3} $= 1.13 \times 10^{-5} h^2 m_p$cm^{-3}

Unit conversions		
1 parsec (pc)	=	3.086×10^{18} cm
1 kilo-parsec (kpc)	=	10^3 pc
1 mega-parsec (Mpc)	=	10^6 pc
1 giga-parsec (Gpc)	=	10^9 pc
1 Astronomical unit (AU)	=	1.5×10^{13} cm
1 year (yr)	=	3.16×10^7 s
1 light year (ly)	=	9.46×10^{17} cm
1 eV	=	1.60×10^{-12} ergs $\equiv 11,604\,\mathrm{K} \times k_B$
1 erg	=	10^{-7} J
Photon wavelength ($\lambda = c/\nu$)	=	1.24×10^{-4} cm (photon energy$/1$ eV)$^{-1}$
1 nano-Jansky (nJy)	=	10^{-32} erg cm^{-2} s^{-1} Hz^{-1}
1 Angstrom (Å)	=	10^{-8}cm
1 micron (μm)	=	10^{-4}cm
1 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.