The First Galaxies in the Universe

Princeton Series in Astrophysics
Edited by David N. Spergel

Theory of Rotating Stars, *by Jean-Louis Tassoul*
Theory of Stellar Pulsation, *by John P. Cox*
Galactic Dynamics, Second Edition, *by James Binney and Scott Tremaine*
Dynamical Evolution of Globular Clusters, *by Lyman Spitzer, Jr.*
Supernovae and Nucleosynthesis: An Investigation of the History of Matter, from the Big Bang to the Present, *by David Arnett*
Unsolved Problems in Astrophysics, *edited by John N. Bahcall and Jeremiah P. Ostriker*
Galactic Astronomy, *by James Binney and Michael Merrifield*
Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment, *by Julian H. Krolik*
Plasma Physics for Astrophysics, *by Russell M. Kulsrud*
Electromagnetic Processes, *by Robert J. Gould*
Conversations on Electric and Magnetic Fields in the Cosmos, *by Eugene N. Parker*
High-Energy Astrophysics, *by Fulvio Melia*
Stellar Spectral Classification, *by Richard O. Gray and Christopher J. Corbally*
Exoplanet Atmospheres: Physical Processes, *by Sara Seager*
Physics of the Interstellar and Intergalactic Medium, *by Bruce T. Draine*
The First Galaxies in the Universe, *by Abraham Loeb and Steven R. Furlanetto*

BB / EA

The First Galaxies in the Universe

Abraham Loeb and
Steven R. Furlanetto

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, 6 Oxford Street,
Woodstock, Oxfordshire OX20 1TW
press.princeton.edu

/ PE

<~?~NOT FINAL; set cipf>
Cover Photograph: *Hubble Extra Deep Field.* Courtesy of NASA, ESA, S. Beckwith (STScI), and the HUDF Team. / EA

~~ISBN 978-0-691-14491-7~~

~~ISBN (pbk.) 978-0-691-14492-4~~

British Library Cataloging-in-Publication Data is available

This book has been composed in Scala Lf

Printed on acid-free paper. ∞

Typeset by S R Nova Pvt Ltd, Bangalore, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

replace ISBNs with insert on following page / EA

Library of Congress Cataloging-in-Publication Data
Loeb, Abraham.
The first galaxies in the universe / Abraham Loeb and Steven R. Furlanetto.
p. cm. -- (Princeton series in astrophysics)
Summary: "This book provides a comprehensive, self-contained introduction to one of the most exciting frontiers in astrophysics today: the quest to understand how the oldest and most distant galaxies in our universe first formed. Until now, most research on this question has been theoretical, but the next few years will bring about a new generation of large telescopes that promise to supply a flood of data about the infant universe during its first billion years after the big bang. This book bridges the gap between theory and observation. It is an invaluable reference for students and researchers on early galaxies. The First Galaxies in the Universe starts from basic physical principles before moving on to more advanced material. Topics include the gravitational growth of structure, the intergalactic medium, the formation and evolution of the first stars and black holes, feedback and galaxy evolution, reionization, 21-cm cosmology, and more. Provides a comprehensive introduction to this exciting frontier in astrophysics Begins from first principles Covers advanced topics such as the first stars and 21-cm cosmology Prepares students for research using the next generation of large telescopes Discusses many open questions to be explored in the coming decade "-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-0-691-14491-7 (hardback) -- ISBN 978-0-691-14492-4 (paper) 1. Galaxies--Formation. 2. Stars--Formation. 3. Cosmology. I. Furlanetto, Steven R. II. Title.
QB857.5.E96L64 2013
523.1'12--dc23
2012018181

To our families

Contents

41 / PE

9 / PE

Preface

This book captures the latest exciting developments concerning one of the most fascinating unsolved mysteries about our origins: *how did the first stars and galaxies form?* This era, known as the *cosmic dawn* because these sources were the first to illuminate our Universe, assumes central importance in our understanding of the history of the Universe. Most research on this question has been theoretical so far. But the next decade or two 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 m in diameter), as well as several new arrays of dipole antennas operating at low radio frequencies. The fresh data on the first galaxies and the diffuse gas 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. Thus 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.

The book is organized into three parts that largely build on each other. Part I, *Fundamentals of Structure Formation*, includes chapters on basic cosmology, linear perturbation theory, nonlinear structure formation, and the intergalactic medium. This part provides a broad introduction to studies of cosmological structure and galaxy formation with applications well beyond the first galaxies themselves. The first three chapters provide a crucial introduction to the rest of the book; the fourth (on the intergalactic medium) is not essential for many of the later chapters but is important for understanding the reionization process as well as many of the most important observational probes of the cosmic dawn.

Part II, *The First Structures*, focuses on the physics driving the formation of these objects, as well as the physics that determines their influence on subsequent generations of objects. We review the formation of the first stars and black holes, the importance of stellar feedback, the basic principles of galaxy evolution, and the epoch of reionization. Many of the principles contained here also have wide application to other areas of extragalactic astrophysics, though we focus on their application to the first galaxies. The first two chapters in this part build on each other, but the others can be approached largely independently.

Part III, *Observations of the Cosmic Dawn*, describes several directions in which we hope to observe the most distant galaxies in the coming decades. This part begins with a discussion of galaxy surveys and then moves on to two unique probes of this era: the Lyman-α and 21-cm lines of neutral hydrogen. It concludes with brief discussions of several other techniques. The chapters in this section are largely independent of each other and may be read in any order.

Throughout the text, we reference seminal papers as well as some recent calculations with endnotes; these are collected in the Notes section. In the Further Reading section, we list useful overviews in the form of books and review papers. We have also included two appendixes. In Appendix A, we include fundamental constants and conversion factors, and, in Appendix B we summarize the cosmological parameters assumed in this book (see also §1.4).

Note that both for the sake of brevity and because the current cosmological measurements are reasonably secure, most of the equations do not explicitly state their dependence on such factors as the baryon density, Hubble constant, or cold dark matter density. Inserting these dependencies is a useful exercise, and we encourage the interested readers to check their understanding in this way.

Various introductory sections of this book are based on an undergraduate-level book, *How Did the First Stars and Galaxies Form?* by one of us (A.L.), which followed a cosmology class he had taught over two decades in the astronomy and physics departments at Harvard University. Other parts relate to overviews 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.

use capital p (Physics)

use capital a (Astronomy)

The writing of this book was made possible thanks to the help we received from many individuals. First and foremost, we are grateful to our families for their support and patience during the lengthy writing period of the book. Needless to say, the content of this book echoes many papers and scientific discussions we had over the years with our students, postdocs, and senior collaborators, including Dan Babich, Rennan Barkana, Jon Bittner, Laura Blecha, Judd Bowman, Frank Briggs, Avery Broderick, Volker Bromm, Chris Carilli, Renyue Cen, Benedetta Ciardi, T. J. Cox, Mark Dijkstra, Daniel Eisenstein, Claude-André Faucher-Giguère, Richard Ellis, Idan Ginsburg, Zoltan Haiman, Lars Hernquist, Jackie Hewitt, Loren Hoffman, Bence Kocsis, Girish Kulkarni, Adam Lidz, Andrei Mesinger, Matt McQuinn, Joey Muñoz, Ramesh Narayan, Peng Oh, Ryan O'Leary, Rosalba Perna, Tony Pan, Ue-Li Pen, Jonathan Pritchard, Fred Rasio, Martin Rees, Doug Rubin, George Rybicki, Athena Stacy, Dan Stark, Yue Shen, Nick Stone, Anne Thoul, Hy Trac, Eli Visbal, Stuart Wyithe, and Matias Zaldarriaga. We did not attempt to provide a comprehensive reference list of the related literature, since such a list would be out of date within a few years in this rapidly evolving frontier. Instead we focused pedagogically on the underlying physical principles that will remain relevant in the future, and referred the reader to representative papers, review articles, and books for further reading. We thank Nina Zonnevylle and Uma Mirani for their assistance in obtaining permissions for the figures of the book; Laurie Lites for her

assistance with the manuscript; Fred Davies, Lauren Holzbauer, Joey Muñoz, and Ramesh Narayan for their help with several figures; and Natalie Mashian, Doug Rubin, and Anjali Tripathi for their comments on the finished manuscript. Finally, it has been a delightful experience for us to work with our book editor, Ingrid Gnerlich, and the entire production team at Princeton University Press.

—A. L. & S. F.

kill back halftitle and blank verso to make even forms / EA

The First Galaxies in the Universe

PART I
Fundamentals of Structure Formation

Chapter One

Introduction and Cosmological Background

1.1 Preliminary Remarks

On large scales, the Universe is observed to be expanding. As it expands, galaxies separate from one another, and the density of matter (averaged over a large volume of space) decreases. If we imagine playing the cosmic movie in reverse and tracing this evolution backward 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 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 is expanding. 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. 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 in which matter was preeminent and led 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 light takes a finite time 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 400,000 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: between these two epochs was a period when

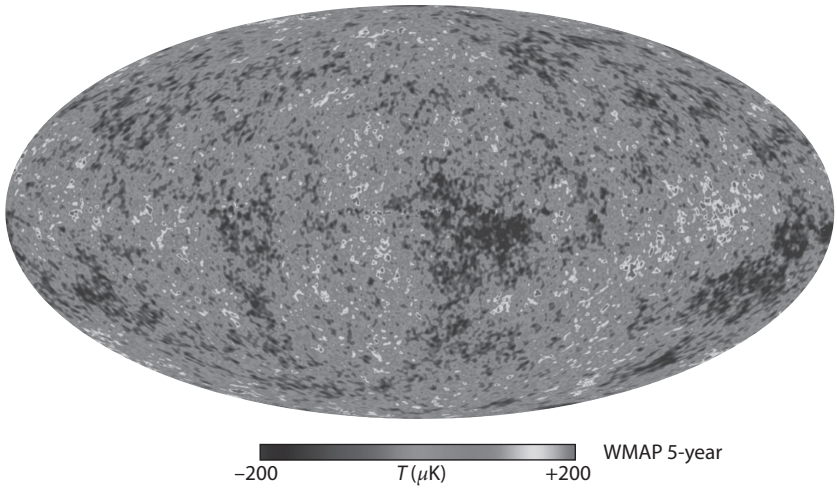


Figure 1.1 Image of the Universe when it first became transparent, 400,000 years after the Big Bang, taken over 5 years by the *Wilkinson Microwave Anisotropy Probe* (WMAP) satellite (see *Color Plate 1* for a color version of this figure). Slight density inhomogeneities at the level of one part in $\sim 10^5$ 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 microkelvins, and the unperturbed temperature is 2.725 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. Courtesy of / NASA and the WMAP Science Team.

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 ultrasound 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 later 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.

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 24–42 m in diameter and NASA's successor to the Hubble Space Telescope, the James Webb Space Telescope (JWST). 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 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 and resembled 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 telescopes described and on providing motivation for these ambitious observational projects.

All these predictions are generated in the context of the modern cosmological paradigm, which turns the Big Bang model into a quantitative tool for understanding our Universe. In the remainder of this chapter, we briefly describe the essential aspects of this paradigm for understanding the formation of the first galaxies in the Universe.

1.2 Standard Cosmological Model

1.2.1 Cosmic Perspective

In 1915 Einstein formulated 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 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). Einstein then applied his equation to describe the global dynamics of the Universe.

There were no computers available in 1915, and Einstein's equations for the Universe were particularly difficult to solve in the most general case. To get around this obstacle Einstein considered 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 seen 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.

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 a convenient "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).

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 "spiral nebulae," which we now know are distant galaxies, as constituents of 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 only after he added a cosmological constant, whose negative gravity would exactly counteract that of matter. However, Einstein later 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 solutions to Einstein's equations for a homogeneous and isotropic Universe. The Universe must be either 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 from us at a speed v that is proportional to their distance r, namely, $v = H_0 r$, where H_0 is a spatial constant (which can evolve with time), commonly termed the *Hubble constant*.[i] Hubble's data indicated that the Universe is expanding. (Hubble also resolved individual bright stars in these nebulae, unambiguously determining their nature and their vast distances from the Milky Way.)

Einstein was remarkably successful in asserting the cosmological principle. As it turns out, our latest data indicate that the real Universe is homogeneous and isotropic on the largest observable scales to within one part in 10^5. In particular, isotropy is well established for the distribution of faint radio sources, optical galaxies, the X-ray background, and most important, the cosmic microwave background (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. 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 easily be solved.

[i] 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 (based on Hubble and Humason's data) was first formulated by Georges Lemaître in 1927, 2 years prior to the observational paper written by Hubble and Humason.

COMP: From the Memo to Compositor:
6) There are footnotes (with roman numerals) as well as endnotes (with Arabic numerals). Follow ms instructions for numbering per chapter.
THE INITIAL FOOTNOTE NUMBERS ARE ALL INCORRECT ("1" rather than "i") AND MUST BE FIXED AS PEs. THE CALLOUTS ARE OK AS SET.

Why was the Universe prepared to be in this special state? Cosmologists were able to go one step further and demonstrated 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 (although other explanations also may create such conditions). 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 for 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 CMB. At early times, this radiation coupled extremely well to the cosmic gas and produced a spectrum known as a *blackbody*, a form 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 calculation gives a unique answer, $\sim r/v \sim 1/H_0$, that is independent of distance.[ii] The latest measurements of the Hubble constant give a value of $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, which implies 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 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

[ii] 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 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).

slightly denser than the mean, we will infer how its density contrast will evolve relative to the background Universe.

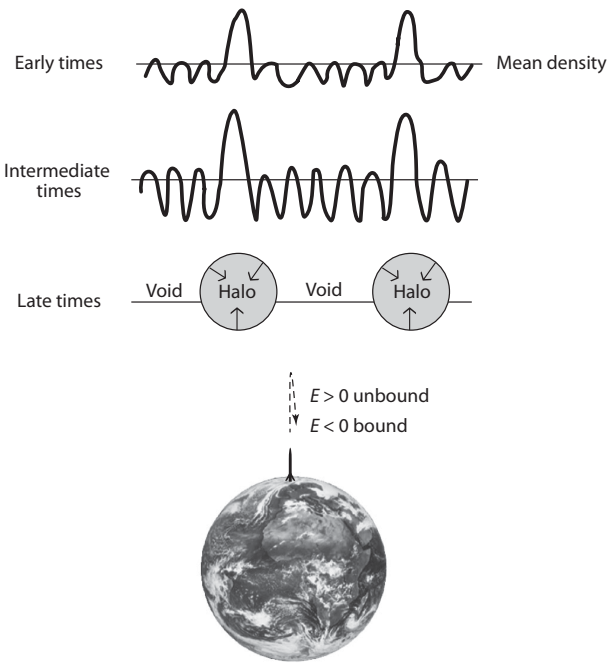
For the moment, let us ignore the energy density of the vacuum (which is always a good approximation at sufficiently early cosmic times, when matter was denser). Then, the equation describing the motion of a spherical shell of matter is identical with the equation of motion of a rocket launched from 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. To deduce 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 has a total mass $M = \rho \times \left(4\pi r^3/3\right)$ enclosed within it. Its energy per unit mass is the sum of the kinetic energy due to its expansion speed $v = Hr$, $(1/2)v^2$, and its potential gravitational energy, $-GM/r$ (where G is Newton's constant), namely, $E = v^2/2 - GM/r$. By substituting the preceding relations for v and M, we can easily show that $E = (1/2)v^2(1 - \Omega)$, 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 toward 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 the 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 in a position to understand how objects like the Milky Way galaxy formed out of small density inhomogeneities that are amplified by gravity.

Let us consider for simplicity the background of a marginally bound (flat) Universe dominated by matter. In such a background, only a slight enhancement in density is required to exceed the critical density ρ_c. Because of Birkhoff's theorem, a spherical region denser than the mean will behave as if it is part of a closed Universe and will increase its density contrast with time, while an underdense spherical region will behave as if it is part of an open Universe and will 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 on themselves (like the trajectory of a rocket launched straight up, away from the center of the earth). An initially slightly inhomogeneous Universe 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 $\sim 10^{-5}$ when the cosmic microwave background last scattered. The overdensities were amplified at late times (once matter dominated



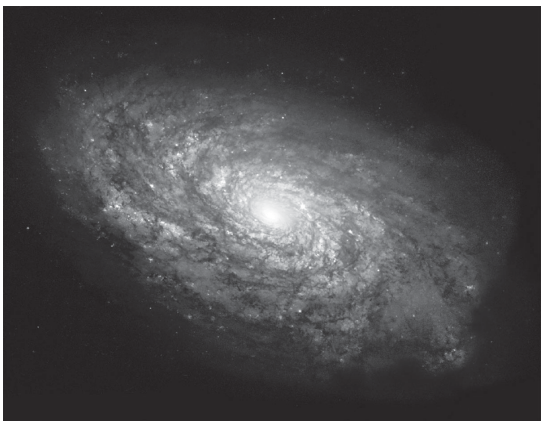


Figure 1.2 *Top:* Schematic illustration of the growth of perturbations to collapsed halos through gravitational instability. The overdense regions initially expand, reach a maximum size, and then turn around and collapse to form gravitationally bound halos if their density exceeds a critical threshold (see §3.1). 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 launched from the surface of the earth depends on the sign of its energy per unit mass, $E = (1/2)v^2 - GM_\oplus/r$. 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 (NASA and ESA). The halo gas cools and condenses to a compact disk surrounded by an extended dark matter halo.

Courtesy of /

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 and the main subject of this book.

Beyond being uniform, the early Universe was simple in one additional way: the process of *Big Bang nucleosynthesis* produced the first atomic nuclei, but these were mostly hydrogen and helium (with ∼93% of those atoms in the form of hydrogen). However, 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 million 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

The history and fate of our Universe are thus tied inexorably to its contents—be it matter, dark energy, or something even more exotic. However, measuring the average density of the Universe is extraordinarily difficult. Fortunately, Einstein's equations show that the contents of the Universe are also tied to its geometry—so measuring the latter would indirectly constrain its components.

How can we tell the difference between the flat surface of a book and the curved surface of a balloon? A simple way is 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° (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°. (This can easily be figured out by examining a globe and noticing that any line connecting one of the poles to the equator opens an angle of 90° 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°.) 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 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 outward. Similarly, perturbing the smooth Universe at a single point at the Big Bang would have resulted in the propogation of a spherical sound wave outward from that point. The wave would have traveled at the speed of sound, which was of the order of the speed of light c (or, more precisely, $c/\sqrt{3}$) 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 yet been able to reach them. 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 will therefore show a characteristic correlation scale, corresponding to the sound horizon at the time when the CMB was produced, 400,000 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 with theory, experimental cosmologists inferred from the simple geometry of triangles that the Universe is flat (or at least very close to it).

The inferred flatness may be a natural consequence of the early period of 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 nonexpanding triangle placed on this patch would get arbitrarily close to 180° as the balloon inflated.

Even though we now know that our Universe is very close to being flat, this flatness only constrains the cumulative energy density in the Universe; it tells us very little about how that energy is distributed among the different components, such as baryons, other forms of matter, and dark energy. We must probe our Universe in other ways to learn about this distribution.

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 is 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 the force of 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; to notice the expansion we need to observe sources spread over the vast scales of millions of light-years.

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

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

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

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

italics dR superscript 2 /PE

where $d\Omega = d\theta^2 + \sin^2\theta\, d\phi^2$, (R, θ, ϕ) are the comoving spherical coordinates centered on the observer, and $(x, y, z) = (R\cos\theta, R\sin\theta\cos\phi, R\sin\theta\sin\phi)$. Throughout this book, we quote distances in these comoving units—as opposed to their time-varying *proper* equivalents—unless otherwise specified.

A source located at a separation $r = a(t)R$ from us will 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 (when $a(t) \equiv 1$). 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 when the car is approaching us than when it is receding from us. 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

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

with the solution $v \propto a^{-1}$. Correspondingly, the wavelength scales as $\lambda = (c/v) \propto a$. 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. This relation holds also for the de Broglie wavelength, $\lambda_{\rm dB} = (h/p) \propto a$, characterizing the quantum-mechanical wavefunction of a massive particle with momentum p (where h is Planck's constant). Consequently, the kinetic energy of a nonrelativistic particle scales as $(p^2/2m_p) \propto a^{-2}$. Thus, in the absence of heat exchange with other systems, the temperature of a gas of nonrelativistic protons and electrons will cool faster ($\propto a^{-2}$) than the CMB temperature ($h\nu \propto a^{-1}$) as the Universe expands and a increases. 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 (that ultimately approaches the speed of light when the redshift goes to infinity), which in turn implies a larger distance (that ultimately approaces our horizon, which is the distance traveled by light since the Big Bang) and an earlier emission time of the source 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).[iii] But there is a limit to how far back we can see: we can image the Universe only if it is transparent. Earlier than 400,000 years after the Big Bang, the cosmic gas was sufficiently hot to be fully ionized, and the Universe was opaque owing 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). 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 as well. 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. There are several examples to consider. For a radiation fluid,[iv] $p_{\rm rad}/c^2 = (1/3)\rho_{\rm rad}$, which implies that $\rho_{\rm grav} = 2\rho_{\rm rad}$.

However, if the vacuum has a nonzero energy density that is constant in space and time, the so-called *cosmological constant*, then the pressure of the vacuum is negative, because by opening up a new volume increment ΔV one *gains* an energy $\rho_{\rm vac}c^2\Delta V$ instead of losing it, 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 relation 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 6 billion years of cosmic history.

AA

[iii] Cosmology and archaeology share another similarity: both are *observational*, rather than *experimental*, sciences. Thus, we are forced to interpret the complicated physics of actual systems rather than design elegant experiments that can answer targeted questions. Although simplified models can be built in the laboratory (or even inside computers), the primary challenge of cosmology is figuring out how to extract useful information from real and complex systems that cannot be artificially altered.

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

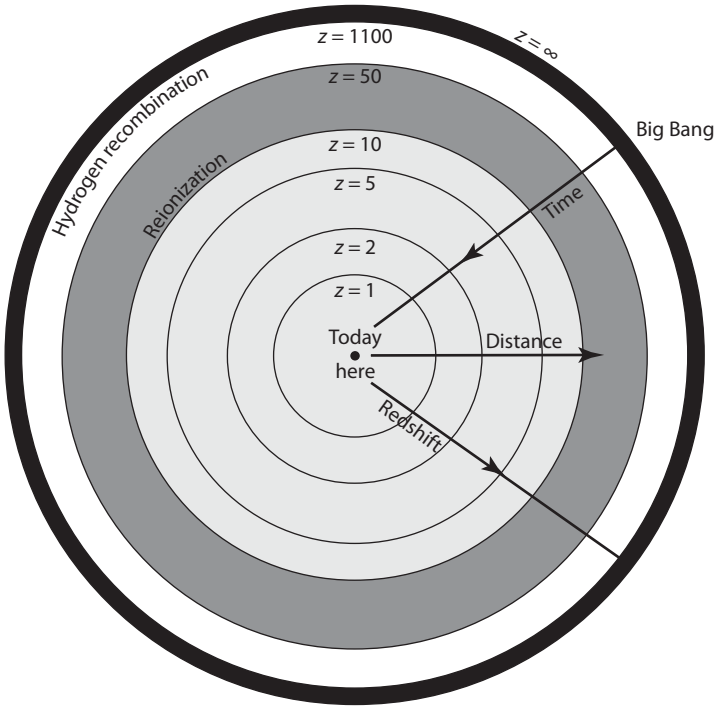


Figure 1.3 Cosmic archaeology of the observable volume of the Universe, in comoving coordinates (which factor out the cosmic expansion). The outermost observable boundary ($z = \infty$) marks the comoving distance that light has traveled since the Big Bang. Future observatories aim to map most of the observable volume of our Universe and to improve dramatically the statistical information we have about the density fluctuations within it. Existing data on the CMB probe 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 dark gray. Note that the comoving volume out to any of these redshifts scales as the distance cubed.

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 decayed when inflation ended).

delete R

change italics r to italics a

should be italics a^2 without subscript r (in the denominator)

In this book, we will primarily be concerned with the *cosmic dawn*, or the era in which the first galaxies formed at $z \sim 6$–30. At these early times, the cosmological constant was very small compared with the matter densities and can generally be ignored.

1.3 Milestones in Cosmic Evolution

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

$$\frac{d^2 r}{dt^2} = -\frac{G M_{\rm grav}}{a_r^2}. \tag{1.4}$$

Since $\rho_{\rm grav} = \rho + 3p/c^2$, we need to know how pressure evolves with the expansion factor $a(t)$. We obtain this information from the thermodynamic relation mentioned previously between the change in the internal energy $d(\rho c^2 V)$ and the $p\,dV$ work done by the pressure, $d(\rho c^2 V) = -p\,dV$. This relation implies $-3pa\dot{a}/c^2 = a^2\dot{\rho} + 3\rho a\dot{a}$, where an overdot 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, $H = \dot{a}/a$, we can rewrite equation (1.5) as

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

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

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

If we denote 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*, with Ω_m, Ω_Λ, and Ω_r, respectively, a flat universe with $E = 0$ 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 (i.e., $\Omega_0 = 1$), then $a \propto t^{2/3}$; if radiation dominates, then $a \propto t^{1/2}$;

and if the vacuum density dominates, then $a \propto \exp\{H_{\rm vac}t\}$, where $H_{\rm vac} = (8\pi G\rho_{\rm vac}/3)^{1/2}$ is a constant. After inflation ended, the mass density of our Universe, ρ, was at first dominated by radiation at redshifts $z > 3{,}300$, by matter at $0.3 < z < 3{,}300$, and finally by the vacuum at $z < 0.3$. The vacuum had already started to dominate $\rho_{\rm grav}$ at $z < 0.7$, or 6 billion years ago. Figure 1.6 illustrates the mass budget in the present-day Universe and during the epoch when the first galaxies formed.

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

$$t \approx \frac{2}{3H_0{\Omega_m}^{1/2}(1+z)^{3/2}} = \frac{0.95\times10^9\,{\rm yr}}{[(1+z)/7]^{3/2}}. \tag{1.9}$$

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

Second, we note the remarkable exponential expansion for a vacuum-dominated phase. This accelerated expansion serves an important purpose in explaining a few puzzling features of our Universe. We have 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 toward us. Therefore, it should take them twice as long to bridge 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 and folded neatly, we know that there must have a been a process that caused this to happen. 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 $\sim$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 was 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. 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 a 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 reentered 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 subatomic quantum-mechanical fluctuations at early times.

After inflation, an unknown process, called "baryogenesis" or "leptogenesis," generated an excess of particles (baryons and leptons) over antiparticles.[v] As the Universe cooled to a temperature of hundreds of millions of electron-volts (where $1\,\mathrm{MeV}/k_B = 1.1604 \times 10^{10}\,\mathrm{K}$), protons and neutrons condensed out of the primordial quark–gluon plasma through the so-called quantum chromodynamics (QCD) phase transition. At about one second after the Big Bang, the temperature declined to $\sim 1\,\mathrm{MeV}$, and the weakly interacting neutrinos decoupled. Shortly afterward, the abundance of neutrons relative to protons froze, and electrons and positrons annihilated one another. 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 $\sim 3{,}000\,\mathrm{K}$, and free electrons recombined with protons to form neutral hydrogen atoms. As soon as the dense fog of free electrons was depleted, the Universe became transparent to the relic radiation, which is observed at present as the CMB. These milestones of the thermal history are depicted in Figure 1.4.

The Big Bang is the only known event in our past history in which particles interacted with center-of-mass energies approaching the so-called Planck scale[vi] $[(hc^5/G)^{1/2} \sim 10^{19}\,\mathrm{GeV}]$, 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.

1.3.1 Luminosity and Angular-Diameter Distances

When we look at our image reflected off a mirror at a distance of 1 m, we see the way we looked 6 nanoseconds ago, the time it took light to travel to the mirror and back. If the mirror is spaced $10^{19}\,\mathrm{cm} = 3\,\mathrm{pc}$ away, we will see the way we looked 21 years ago. Light propagates at a finite speed, so by observing distant regions, we are able to see what 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

[v] The origin of the asymmetry in the cosmic abundance of matter over antimatter is still an unresolved puzzle.

[vi] 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.

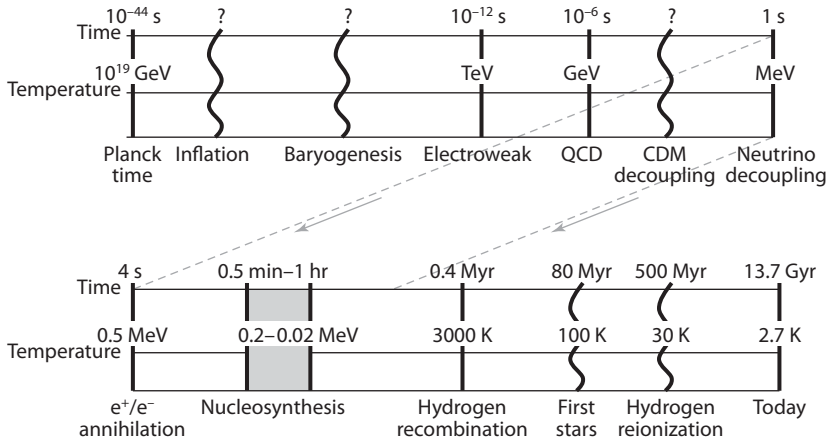


Figure 1.4 Following inflation, the Universe went through several other milestones that left a detectable record. These include baryogenesis (which resulted in the observed asymmetry between matter and antimatter), the electroweak phase transition (during which the symmetry between electromagnetic and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

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.

How faint will the earliest galaxies appear to our telescopes? In an expanding Universe there is some ambiguity as to which "distance" is most relevant. For example, the framework we described earlier—in which the clocks are synchronized relative to the Big Bang—is not appropriate for observations, because light has a finite speed, so that a signal emitted from one clock at time t_A will be observed by another clock at a time $t_B > t_A$. Which of these times should we use to compute the scale factor in a distance formula? Moreover, the *method* of observation influences the choice of the relevant distance, because the photons themselves evolve as they travel.

To answer these questions, 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 the effective distance of the source,

$$f = \frac{L}{4\pi d_{\rm L}^2}, \tag{1.10}$$

where $d_{\rm L}$ is defined as the *luminosity distance* in cosmology. For a flat Universe, the comoving distance of a galaxy that emitted its photons at a time $t_{\rm em}$ and is observed at time $t_{\rm obs}$ is obtained by summing over infinitesimal distance elements along the path length of a photon, $c\,dt$, each expanded by a factor $(1+z)$ to the present time (corresponding to setting $ds^2 = 0$ in equation 1.1 for a photon trajectory):

$$R_{\rm em} = \int_{t_{\rm em}}^{t_{\rm obs}} \frac{c\,dt}{a(t)} = \frac{c}{H_0}\int_0^z \frac{dz'}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}, \tag{1.11}$$

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

If a galaxy has an intrinsic luminosity L, then it will emit an energy $L\,dt_{\rm em}$ 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 will be

$$f = \frac{L\,dt_{\rm em}/(1+z)}{4\pi R_{\rm em}^2 dt_{\rm obs}} = \frac{L}{4\pi R_{\rm em}^2(1+z)^2}, \tag{1.12}$$

which implies that

$$d_{\rm L} = R_{\rm em}(1+z) = d_{\rm A}(1+z)^2. \tag{1.13}$$

Unfortunately, for a flat universe with a cosmological constant, these distance integrals cannot be expressed analytically. However, a convenient numerical approximation, valid to 0.4% relative error in the range $0.2 \le \Omega_m \le 1$ (where Ω_m is the total matter density) is[1]

$$d_L = \frac{c}{H_0} a^{-1}\left[\eta(1,\Omega_m) - \eta(a,\Omega_m)\right], \tag{1.14}$$

where

$$\eta(a,\Omega_m) = 2\sqrt{s^3+1}\left[\frac{1}{a^4} - 0.1540\frac{s}{a^3} + 0.4304\frac{s^2}{a^2} + 0.19097\frac{s^3}{a} + 0.066941 s^4\right]^{-1/8}, \tag{1.15}$$

and $s^3 = 1/\Omega_m - 1$.

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

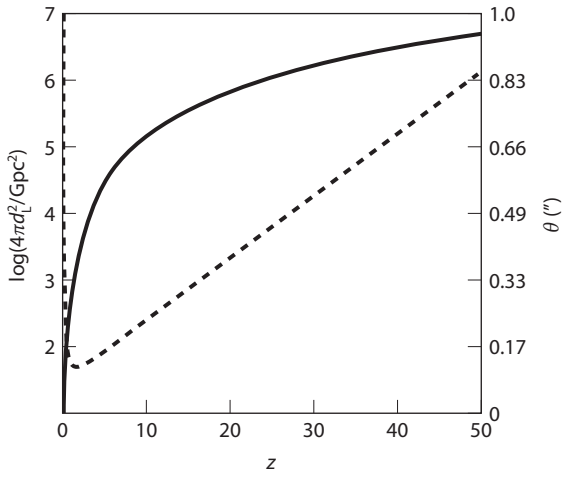


Figure 1.5 The solid line (corresponding to the label on the left-hand side) shows $\log_{10}$ of the conversion factor between the luminosity of a source and its observed flux, $4\pi d_L^2$ (in Gpc^2), as a function of redshift, z. The dashed-dotted line (labeled on the right) gives the angle θ (in arcseconds) occupied by a galaxy of 1 kpc diameter as a function of redshift.

The area dilution factor $4\pi d_L^2$ is plotted as a function of redshift in the solid curve of Figure 1.5. If the observed flux is measured over only a narrow band of frequencies, one needs to take account of the additional conversion factor $(1+z) = (d\nu_{em}/d\nu_{obs})$ between the emitted frequency interval $d\nu_{em}$ and its observed value $d\nu_{obs}$. This correction yields the relation $(df/d\nu_{obs}) = (1+z) \times (dL/d\nu_{em})/(4\pi d_L^2)$.

In practice, observed brightnesses are often expressed using the *AB magnitude* system. The conversion from flux density to AB magnitude is

$$\mathrm{AB} = -2.5 \log_{10}\left[\frac{df}{d\nu_{obs}}\right] - 48.6, \qquad (1.16)$$

where the flux density is expressed in units of erg s^{-1} cm^{-2} Hz^{-1}.

1.4 Most Matter is Dark

Surprisingly, most of the matter in the Universe is not the same ordinary matter of which we are made (see Figure 1.6). If it were ordinary matter (which also makes stars and diffuse gas), it would have interacted with light, thereby revealing its existence to observations through telescopes. Instead, observations of many different astrophysical environments require the existence of some mysterious dark component of matter that reveals itself only through its gravitational influence and leaves no other clue about its nature. Cosmologists are like detectives who find evidence for some unknown criminal at a crime scene

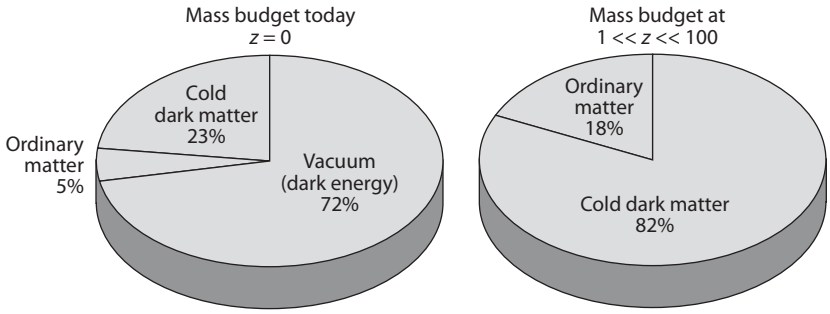


Figure 1.6 Mass budgets of different components in the present-day Universe and in the infant Universe when the first galaxies formed (redshifts $z = 10$–50). The CMB radiation (not shown) makes up a fraction ($\sim$0.03%) of the budget today but was dominant at redshifts $z > 3{,}300$. The cosmological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

and are anxious to find his or 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 never have existed by now, because ordinary matter is coupled to the CMB radiation that filled the early Universe. 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 were 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 never have 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 remembering 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 which the Sun was made as one out of tens of billions of stars, and Earth was born out of the debris left over from the formation process of the Sun. This sequence of events would never have occurred without the dark matter.

We do not know what constitutes the dark matter, but from the good match obtained between observations of large-scale structure and the equations describing a pressureless fluid (see equations 2.3–2.4), we infer that it is likely made of particles with small random velocities. It is therefore called "cold dark matter" (CDM). The popular view is that CDM is composed of particles that weakly interact with ordinary matter, much like the elusive neutrinos we know

to exist. The abundance of such particles would naturally "freeze out" at a temperature $T > 1\,\mathrm{MeV}$, at which the Hubble expansion rate is comparable to the annihilation rate of the CDM particles. Interestingly, such a decoupling temperature, together with a weak interaction cross section and particle masses of $mc^2 > 100\,\mathrm{GeV}$ (as expected for the lightest, and hence stable, supersymmetric particle in simple extensions of the standard model of particle physics), naturally leads through a Boltzmann suppression factor $\sim \exp(-mc^2/k_BT)$ to $\Omega_m \sim 1$. The hope is that CDM particles, owing to their weak but nonvanishing 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). 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's being successful in finding a DNA sample of the previously unidentified criminal.

The most popular candidate for the 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 crosssection for scattering off standard model particles sets the epoch of their thermal decoupling from the cosmic plasma.

In addition to dark matter, the observed acceleration in the current expansion rate of the Universe implies that the vacuum contributes $\sim$72% of the cosmic mass density at present. If the vacuum density will behave as a cosmological constant, it will dominate even more in the future (since $\rho_m/\rho_v \propto a^{-3}$). The exponential future expansion will carry all galaxies outside the local group out of our horizon within $\sim 10^{11}$ years,[2] and will stretch the characteristic wavelength of the cosmic microwave background to be larger than the horizon in $\sim 10^{12}$ years.[3]

The dark ingredients of the Universe can be probed only 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 2.

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

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

collapsed owing to their enhanced gravity and eventually formed gravitationally bound halos, first on small spatial scales and later on larger ones. In this hierarchical model of galaxy formation, the small galaxies formed first and then merged, or accreted gas, to form larger galaxies. At each snapshot of this cosmic evolution, the abundance of collapsed halos, whose masses are dominated by dark matter, can be computed from the initial conditions. The common understanding of galaxy formation is based on the notion that stars formed out of the gas that cooled and subsequently condensed to high densities in the cores of some of these halos.

Gravity thus explains how some gas is pulled into the deep potential wells within dark matter halos and forms galaxies. One might naively expect that the gas outside halos would remain mostly undisturbed. However, observations show that it did not remain 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. This *reionization epoch* marks a crucial transition in the history of the Universe and is a prime foucs of both modern astrophysics research and this book.

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 (see also Appendix B). However, thousands of books in libraries throughout the world cannot summarize the complexities of galaxies, stars, planets, life, and intelligent life in the present-day Universe. If we feed the simple initial cosmic conditions into a gigantic computer simulation incorporating the known laws of physics, we should be able to reproduce all the complexity that emerged out of the simple early Universe. Hence, all the information associated with this later complexity was encapsulated in those simple initial conditions. We will follow the process through which late-time complexity appeared and established an irreversible arrow to the flow of cosmic time.[viii]

[viii] In previous decades, astronomers used to associate the simplicity of the early Universe with the fact that the data about it were 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.

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*? 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 has so far remained veiled in older galaxies by complex astrophysical processes.

Chapter Two

Linear Growth of Cosmological Perturbations

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 until they eventually collapsed to make the first galaxies. Here we describe the first stages of that process and introduce the methods conventionally used to describe these fluctuations.

2.1 Growth of Linear Perturbations

As discussed earlier, small perturbations in density grow owing 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 $\Omega_m \approx 1$. Given that the mean cosmic density was close to the threshold for collapse, a spherical region that 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, 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 (or proper) and comoving coordinates (the latter expand with the background

Universe). In 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

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

where the mean fluid mass density is $\bar{\rho}$, with a corresponding peculiar velocity that 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, where the bulk velocity vanishes, we have

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

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

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

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

The pressure p depends on the species under consideration. For cold dark matter, it vanishes; for an ideal gas of baryons at a fixed temperature, the pressure perturbation is $(\delta p) = c_s^2\delta\bar{\rho}$. The sound speed for a monatomic gas that obeys the ideal gas equation of state $p = nkT_e$ and undergoes Hubble expansion is

$$c_s^2 = \frac{dp/da}{d\rho/da} = \frac{k_B T_e}{\mu m_p}\left(1 - \frac{1}{3}\frac{d\log T_e}{d\log a}\right), \tag{2.5}$$

lc / PE

where T_e is the gas kinetic temperature, and μ is the mean molecular mass in units of m_p. (For primordial neutral gas including a mass fraction $Y_p = 0.24$ of helium, $\mu = 1.22$.) In this section we adopt this expression for the sound speed, though we note that it assumes that the temperature traces the density field (see §2.2.1 for a more exact treatment).

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 nonlinear with $|\delta| > 1$, and at that point the individual particle trajectories must in general be followed.

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

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta - \frac{c_s^2k^2}{a^2}\delta, \tag{2.6}$$

where the last term is the pressure force, which vanishes for cold dark matter. In general, this linear equation has two independent solutions, only one of which grows in time. From random initial conditions this "growing mode"

comes to dominate the density evolution. Thus, until it becomes nonlinear, the density perturbation maintains its shape in comoving coordinates and grows in amplitude in proportion to a growth factor $D(t)$. The growth factor in a flat (matter-dominated) Universe at $z < 10^3$ is given by [1]

$$D(t) \propto \frac{\left(\Omega_\Lambda a^3 + \Omega_m\right)^{1/2}}{a^{3/2}} \int_0^a \frac{a'^{3/2}\, da'}{\left(\Omega_\Lambda a'^3 + \Omega_m\right)^{3/2}} \,. \tag{2.7}$$

In the matter-dominated regime of the redshift range $1 < z < 10^3$, the growth factor is simply proportional to the scale factor $a(t)$. The normalization is usually chosen to be relative to the perturbation amplitude at the present day; we will discuss how to determine this factor §2.1.3.

In a flat Universe with a cosmological constant, this integral cannot be written in closed form without special functions. However, an approximation accurate to ∼2% in the range $\Omega_0 > 0.1$ is $D(z) = \mathcal{D}(z)/(1+z)$ with[2]

$$\mathcal{D}(z) = \frac{5\Omega_0(z)}{2}\left[\Omega_0(z)^{4/7} - \Omega_\Lambda(z) + (1+\Omega_0(z)/2)(1+\Omega_\Lambda/70)\right]^{-1}, \tag{2.8}$$

where (if $\Omega_m + \Omega_\Lambda = 1$)

$$\Omega_m(z) = \frac{\Omega_0(1+z)^3}{\Omega_0(1+z)^3 + \Omega_{\Lambda,0}}, \tag{2.9}$$

$$\Omega_\Lambda(z) = \frac{\Omega_{\Lambda,0}}{\Omega_0(1+z)^3 + \Omega_{\Lambda,0}}, \tag{2.10}$$

and $\Omega_{\Lambda,0}$ is the present-day energy density in a cosmological constant scaled to the critical density. Here $\mathcal{D}(z)$ is normalized to equal unity in a matter-dominated Universe. At the high redshifts of most interest to us, this is a reasonable approximation.

Interestingly, in this matter-dominated regime the gravitational potential $\phi \propto \delta/a$ does not grow in comoving coordinates, which implies that the potential depth of fluctuations remains frozen in amplitude as fossil relics from the inflationary epoch during which they were generated. Nonlinear collapse changes the potential depth only by a factor of the order of 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 the density fluctuations freeze in amplitude [$D(t) \to$ constant] as their growth is suppressed by the accelerated expansion of space.

It is usually convenient to express the density field as a sum over a complete set of periodic Fourier modes, each with a sinusoidal (wavelike) dependence on space with a comoving wavelength $\lambda = 2\pi/k$ and wavenumber k. Mathemati-

change subscript from 0 to m \AA

(z)

eliminate ,0 in subscript \AA

uppercase Greek Lambda / PE

cally, we write[i]

$$\delta_{\mathbf{k}} = \int d^3x\, \delta(x) e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{2.11}$$

$$\delta(\mathbf{x}) = \int \frac{d^3k}{(2\pi)^3} \delta_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{2.12}$$

where $\mathbf{x}$ is the comoving spatial coordinate. The characteristic amplitude of each $\mathbf{k}$-mode defines the typical value of δ on the spatial scale λ. It is straightforward to show that equation (2.6) applies to each Fourier mode individually, so the factor $D(t)$ also describes their growth (in the linear regime), and the evolution of the density field in Fourier space is easy to follow. In particular, note that different spatial scales evolve *independently* in the linear regime.

It is also useful to consider the velocity field $\mathbf{u}$. To linear order, the continuity equation (2.2) becomes $\nabla \cdot \mathbf{u} = -a(d\delta/dt)$, or in Fourier space

$$-i\mathbf{k} \cdot \mathbf{u}_{\mathbf{k}} = -\frac{a}{D}\frac{dD}{dt}\delta_{\mathbf{k}}, \tag{2.13}$$

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

$$\mathbf{u}_{\mathbf{k}} = -i\frac{aHf(\Omega)}{k}\delta_{\mathbf{k}}\hat{\mathbf{k}}, \tag{2.14}$$

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

2.1.1 The Power Spectrum of Density Fluctuations

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

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

$$\xi(\mathbf{x}) = \langle \delta(\mathbf{x})\delta(0) \rangle\,, \tag{2.15}$$

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

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

The second measure is the *power spectrum*, $P(\mathbf{k})$, defined by

$$P(\mathbf{k}) = \left\langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \right\rangle = (2\pi)^3 \delta^D(\mathbf{k} - \mathbf{k}') P(\mathbf{k}), \tag{2.16}$$

which has units of volume. This is simply related to the variance of the amplitude of waves on a given scale. Again, it is a function only of $k = |\mathbf{k}|$ for an isotropic universe.

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

$$\xi(\mathbf{x}) = \left\langle \int \frac{d^3k}{(2\pi)^3} \delta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \int \frac{d^3k'}{(2\pi)^3} \delta^*_{\mathbf{k}'} \right\rangle \tag{2.17}$$

$$= \int \frac{d^3k}{(2\pi)^3} \int \frac{d^3k'}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} \left\langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \right\rangle \tag{2.18}$$

$$= \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} P(\mathbf{k}), \tag{2.19}$$

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

Inflation generates perturbations in which different $\mathbf{k}$-modes are statistically independent, and each has a random phase constant in its sinusoid. This makes the density field following inflation a *Gaussian random field*, and its statistical properties are perfectly described by the power spectrum (see §2.1.3). In other words, all higher-order moments and correlations are simply functions of the power spectrum (or correlation function): no additional parameters are needed to understand the distribution, at least until nonlinear evolution becomes important (which does induce higher-order correlations purely through gravitational instability). A very small amount of primordial non-Gaussianity can be accommodated by existing observations; the nonlinear phase of gravitational collapse generates more.

Moreover, in the standard cosmological model, inflation produces a very simple primordial power-law spectrum $P(k) \propto k^{n_s}$ with $n_s \approx 1$. Quantum fluctuations during cosmic inflation naturally result in a nearly scale-invariant spectrum because of the near constancy of the Hubble parameter for a nearly steady vacuum density. This spectrum has the special property that gravitational potential fluctuations of all wavelengths have the same amplitude at the time when they enter the horizon (namely, when their wavelength matches the

distance traveled by light during the age of the Universe), so this spectrum is called *scale invariant*. This is easy to see: the mean square amplitude of mass fluctuations within spheres of comoving radius ℓ is $(\delta M/M)^2 \propto k^3 P(k)$ for $k \sim 2\pi/\ell$. Therefore, the corresponding fluctuation amplitude of the gravitational potential, $\sim (G\delta M/\ell) \propto \ell^{(1-n_s)/2}$, is independent of scale if $n_s = 1$. This spectrum has the aesthetic appeal that perturbations can always be small on the horizon scale. A different power-law spectrum would lead to an overdensity of the order of unity across the horizon, and result in black hole formation, either in the future or past of the Universe.

However, the power spectrum becomes more complex as perturbations grow at later times in a CDM universe. In particular, the modified final power spectrum is characterized by a turnover at a scale on the order of 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 at that time ($\sim ct/\sqrt{3}$; see the next chapter) is comparable to the scale of the horizon, inside of which growth is enabled by causality. Therefore, modes on a spatial scale that entered the horizon during the early radiation-dominated era got trapped at their initial small density contrast and so show a smaller amplitude relative to the power-law extrapolation of long-wavelength modes that entered the horizon during the matter-dominated era.

For a scale-invariant index $n_s \approx 1$, the small-scale fluctuations have the same amplitude at horizon crossing, and with nearly no growth they have the same amplitude on all subhorizon mass scales at matter-radiation equality. The associated constancy of the fluctuation amplitude on small mass scales (in real space), $\delta^2 \propto P(k)k^3 \sim$ constant, implies a small-scale asymptotic slope for $P(k)$ of ≈ -3. The resulting power spectrum after matter-radiation equality is often parameterized by a *transfer function* that accounts for changes in the shape of the dark matter power spectrum up to this point. The transfer function is defined so that

$$P(k,z) = T^2(k)\frac{D^2(z)}{D^2(z_{\rm eq})}P_{\rm pri}(k), \tag{2.20}$$

where $P_{\rm pri}(k)$ is the primordial power spectrum. Note that the transfer function is time independent (but scale dependent) because it describes all the evolution from inflation through the era of matter-radiation equality. The growth factor, however is scale independent (but time dependent) because dark matter perturbations do not have any scale dependence during the matter era. The transfer function is crudely described by the fitting function[3]

$$T^2(k)P_{\rm pri}(k) \propto k^{n_s}/(1+\alpha_p k+\beta_p k^2)^2, \tag{2.21}$$

with $\alpha_p = 8(\Omega_m h^2)^{-1}$ Mpc and $\beta_p = 4.7(\Omega_m h^2)^{-2}$ Mpc^2. This function provides a reasonable fit to the overall shape of the power spectrum, but small-scale features and subtle modifications not captured by this simple formula are extremely important as well. These include the effects of neutrinos with finite mass (which wash out small-scale structure, thanks to the relativistic motions of

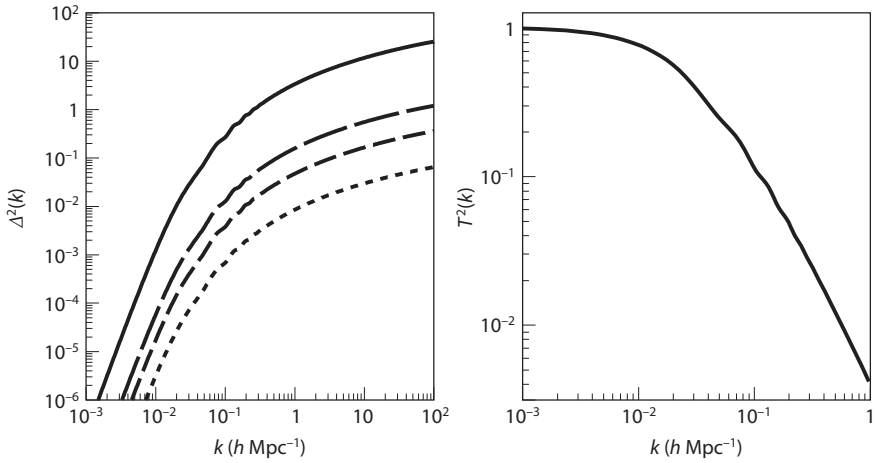


Figure 2.1 *Left:* The matter power spectrum in our fiducial cosmology at $z = 0$, 5, 10, and 25, from top to bottom. *Right:* The corresponding transfer function. Computed using the publicly available code CAMB (http://camb.info).

these particles) and the influence of baryons, which we discuss in detail next.[4] Figure 2.1 shows the resulting matter power spectra and transfer functions at $z = 0$, 5, 10, and 25, using our fiducial cosmological parameters. Note the oscillatory features near $k \sim 0.1h$ Mpc^{-1}, which are called *baryon acoustic oscillations*, whose source we discuss next.

2.1.2 Relative Streaming of Baryons and Cold Dark Matter

Species that decouple at a particular time from the cosmic plasma (including the dark matter and the baryons) will show fossil evidence for acoustic oscillations in their power spectrum of inhomogeneities owing 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 (see also the discussion in §1.2.3;): in the radiation fluid, where $c_s \approx c/\sqrt{3}$, this sound horizon will be near the causal horizon as well. 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 result is a series of peaks in the power spectrum corresponding to the harmonics of this physical scale, as shown in Figure 2.1.

These peaks from radiation coupling to the dark matter sector are on very small spatial scales (for weakly interacting particles, they correspond to mass scales of planets or smaller).[5] The mass scales of the perturbations that grew to become the first collapsed objects at $z < 100$ crossed the horizon in the radiation-dominated era after the dark matter had already decoupled from the cosmic plasma and so were largely unaffected by this streaming.

However, prior to cosmological recombination, the baryons and the cosmic background radiation were tightly coupled and behaved as a single fluid, separate from the dark matter. Because this was relatively late in the history of star formation, the physical scales of these correlations are reasonably large: ∼150 Mpc today. These large-scale features can be incorporated into the transfer function and, because their locations can be predicted from first principles for a given cosmological model, act as "standard rulers" that are useful in measuring the fundamental parameters of our Universe. The induced correlations occur on such large scales that they do not themselves appreciably affect structure formation at high redshifts.

However, a related effect is potentially very important.[6] When the gas decoupled from the radiation at $z \approx 10^3$, it was streaming relative to the dark matter with a root-mean-square (rms) speed of $v_{\rm bc} \approx 10^{-4}c = 30$ km s^{-1}. This speed is much larger than the sound speed, so it has important implications for the accretion of gas onto dark matter structures (see §3.2.2). Here we will show how these effects can be incorporated into perturbation theory to describe gravitational instability.

Using the continuity equation for the baryons and cold dark matter separately, we can write the Fourier transform of the relative velocity between the two species (to linear order) as

$$\mathbf{u}_{bc}(\mathbf{k}) = \frac{\mathbf{k}}{ik^2}[\theta_b(\mathbf{k}) - \theta_c(\mathbf{k})], \tag{2.22}$$

where $\theta \equiv a^{-1}\nabla \cdot \mathbf{u}$. From equation (2.14), the power spectrum of this relative velocity is then

$$\Delta^2_{\rm vbc}(k) = \frac{k^3}{2\pi^2} P_{\rm pri}(k) \left[\frac{\theta_b(k) - \theta_c(k)}{k}\right]^2, \tag{2.23}$$

and the total variance is $\langle u^2_{\rm bc}(\mathbf{x})\rangle = \int (dk/k)\Delta^2_{\rm vbc}(k)$. Figure 2.2 shows the variance of the velocity difference perturbations (in units of c) per $\ln k$ as a function of the mode wavenumber k at $z = 10^3$. The power extends to scales as large as the sound horizon at recombination, $\sim$ 140 Mpc, but declines rapidly at $k > 0.5$ Mpc^{-1}, which indicates that the velocity of the baryons relative to the dark matter was coherent over the photon diffusion (or Silk damping) scale of several comoving megaparsecs. This scale is larger by two orders of magnitude than the size of the regions out of which the first galaxies were assembled at later times. Therefore, in the restframe of those galaxies, the background intergalactic baryons appeared to be moving coherently as a wind. In the next chapter, we will examine whether this wind had a significant effect on the assembly of baryons onto the earliest galaxies.

In the presence of this relative motion between baryons and cold dark matter, the perturbation analysis becomes somewhat more complex. The simplest approach is to treat the two species as having a spatially constant bulk velocity $\mathbf{v}_{\rm bc}$ that decays with redshift as $1/a$ as the neutral gas falls into the gravitational potential wells of the dark matter (see equation 2.22). The assumption of a spatially constant background velocity is valid on scales smaller than the coherence

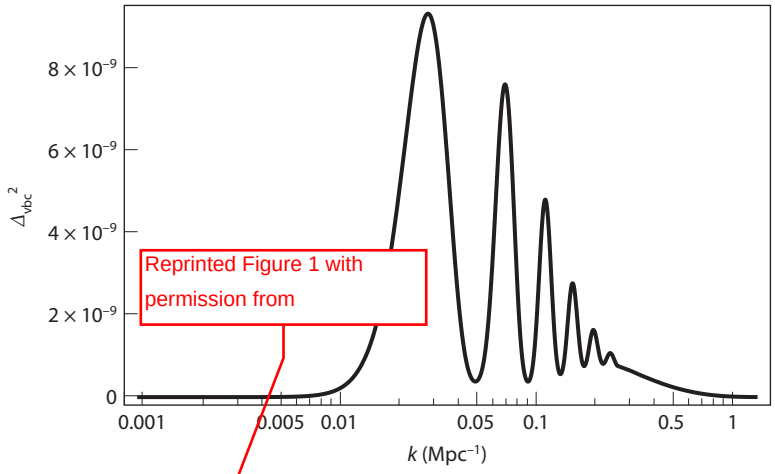


Figure 2.2 The variance of the velocity difference perturbations (in units of c) between baryons and dark matter per $\ln k$ as a function of comoving wavenumber k at $z = 10^3$. Tseliakhovich, D., & Hirata, C., *Phys. Rev.* **D82**, 3520 (2010). Copyright 2010 by the American Physical Society.

length of the velocity field (i.e., several megaparsecs). In the rest frame of the baryons, the analogs to equation (2.2) and (2.3) are (note that we require seperate equations for each type of matter)

$$\frac{\partial \delta_c}{\partial t} = \frac{i}{a}\mathbf{u}_{\rm bc}\cdot\mathbf{k}\delta_c - \theta_c, \tag{2.24}$$

$$\frac{\partial \theta_c}{\partial t} = \frac{i}{a}\mathbf{u}_{\rm bc}\cdot\mathbf{k}\theta_c - \frac{3H^2}{2}(\Omega_c\delta_c + \Omega_b\delta_b) - 2H\theta_c, \tag{2.25}$$

$$\frac{\partial \delta_b}{\partial t} = -\theta_b, \tag{2.26}$$

$$\frac{\partial \theta_b}{\partial t} = -\frac{3H^2}{2}(\Omega_c\delta_c + \Omega_b\delta_b) - 2H\theta_c + \frac{c_s^2k^2}{a^2}\delta_b. \tag{2.27}$$

The first terms on the right-hand side of the cold dark matter equations remain here because the bulk velocity is large and so cannot be ignored during the linearization of the basic fluid equations. When they are large compared with the velocity divergence term, the relative streaming will have a significant effect on structure formation. This occurs at a scale

$$k_{\rm ubc} \sim \frac{aH}{\langle u_{\rm bc}^2\rangle^{1/2}} \sim 180\left(\frac{30\ {\rm km\ s^{-1}}}{\langle u_{\rm bc}^2(z_{\rm rec})\rangle^{1/2}}\right)\left(\frac{1+z}{50}\right)^{-1/2}\ {\rm Mpc}^{-1}, \tag{2.28}$$

where we have scaled to the typical bulk velocity at recombination and used $\mathbf{u}_{bc} \propto (1+z)^{-1}$. The suppression scale is larger at higher redshift, which means that the acoustic feature will affect structure formation to some degree at even

Set Roman / AA

Delete hyphen and close up / PE

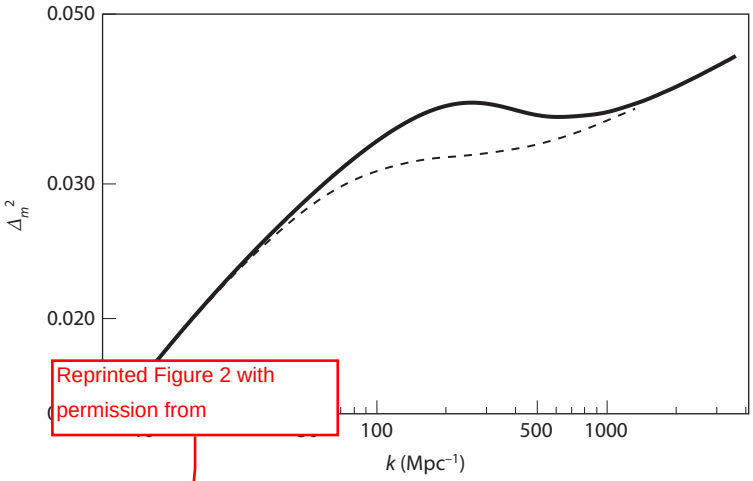


Figure 2.3 The isotropically averaged power spectrum of the matter distribution at $z = 40$ with and without the relative streaming motions (dashed and solid lines, respectively). Tseliakhovich, D., & Hirata, C., *Phys. Rev.* **D82**, 3520 (2010). Copyright 2010 by the American Physical Society.

larger scales than this estimate shows (see also §3.2). Note as well that the relative velocity term is much, much smaller than the divergence term on scales larger than the coherence length, which has $k \sim 1\ \mathrm{Mpc}^{-1}$, so this system of equations is reasonably accurate on large scales as well.

Figure 2.3 shows the effect of these velocities on the total matter power spectrum at high redshifts: because the baryons constitute ∼18% of the matter, the dark matter power spectrum changes significantly on the relevant scales.

2.1.3 Normalizing the Power Spectrum

Although the shape of the power spectrum is well determined by linear perturbation theory in an expanding universe, the overall *amplitude* of the power spectrum is not specified by current models of inflation and is usually set by comparision with the observed CMB temperature fluctuations or with measures of large-scale structure based on surveys of galaxies, clusters of galaxies, or the intergalactic gas.

The most popular large-scale structure normalization is through the observed mass fluctuation amplitude (at the present day) on $8h^{-1}$ Mpc, roughly the scale of galaxy clusters. To relate this quantity to the power spectrum, we must consider the statistical distribution of the smoothed density field. We define a window (or filter) function $W(\mathbf{r})$ normalized so that $\int d^3r\, W(\mathbf{r}) = 1$, where the smoothed density perturbation field is $\int d^3r \delta(\mathbf{x}) W(\mathbf{r})$. The simplest observed quantity is a measure of the masses (relative to the mean) inside spheres of radius R; in this case we use a "spherical top-hat" window (similar to a

three-dimensional cookie cutter), in which W = constant inside a sphere of radius R, and $W = 0$ outside.

The normalization of the present day power spectrum at $z = 0$ is then specified by the variance of this density field when smoothed on the particular scale of $8h^{-1}$Mpc, $\sigma_8 \equiv \sigma(R = 8h^{-1}\text{Mpc})$. For the top-hat filter, the smoothed perturbation field is denoted by δ_R or δ_M, where the enclosed mass M is related to the comoving radius R by $M = 4\pi\rho_m R^3/3$, in terms of the current mean density of matter ρ_m. We then write the variance $\langle\delta_M^2\rangle$ (relative to the mean) as[ii]

$$\sigma^2(M) = \left\langle \frac{1}{V}\int d^3x\,\delta(\mathbf{x})W(\mathbf{x})\,\frac{1}{V}\int d^3x'\,\delta(\mathbf{x}')W(\mathbf{x}')\right\rangle \tag{2.29}$$

$$= \frac{1}{V^2}\int d^3x\,d^3x'\,W(\mathbf{x})W(\mathbf{x}')\xi(|\mathbf{x}-\mathbf{x}'|) \tag{2.30}$$

$$= \int \frac{d^3k}{(2\pi)^3}P(k)\frac{|W_\mathbf{k}|^2}{V^2}, \tag{2.31}$$

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

$$\sigma^2(M) \equiv \sigma^2(R) = \int_0^\infty \frac{dk}{k}\,\Delta^2(k)\left[\frac{3j_1(kR)}{kR}\right]^2, \tag{2.32}$$

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

While the normalization of the power spectrum requires only σ_8, we will see in the next chapter that the function $\sigma(M)$ plays a major role in fixing the abundance of collapsed objects. We therefore show it in Figure 2.4 as a function of mass and redshift for our standard cosmological model. Note that $\sigma^2 \propto \delta^2 \propto D(t)^2$, so the time dependence is trivial (at least in linear theory).

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

$$p(\delta)d\delta = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\delta^2/2\sigma^2}d\delta. \tag{2.33}$$

These so-called Gaussian perturbations are a key prediction of inflation; they have the convenient property that the statistical distribution of densities is described entirely by the power spectrum (through σ^2).

2.2 The Thermal History During the Dark Ages

In addition to the density evolution, the second key "initial condition" for galaxy formation is the temperature of the hydrogen and helium gas that will collapse

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

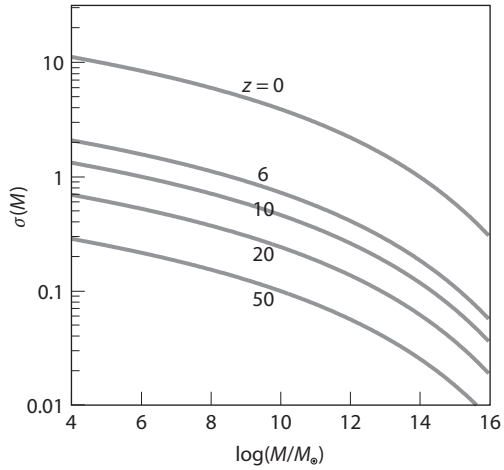


Figure 2.4 The root-mean-square amplitude of linearly extrapolated density fluctuations σ as a function of mass M (in solar masses $M_\odot$, within a spherical top-hat filter) at different redshifts z. Halos form in regions that exceed the background density by a factor of the order of unity. This threshold is surpassed only 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 $R = 1.85(M/10^{12}M_\odot)^{1/3}$ Mpc.

into the first galaxies. If it were isolated, the gas would simply cool adiabatically with the overall expansion of the universe. In general, for an ideal gas this cooling rate can be written as $(\gamma - 1)(\dot{\rho}_b/\rho_b)T_e$, where ρ_b is the baryon density, and $\gamma = 5/3$ is the adiabatic index of a monatomic gas. For gas at the mean density, the factor $(\dot{\rho}_b/\rho_b) = -3H$ owing to the Hubble expansion.

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

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

$$T(\theta) = T_\gamma \left(1 + \frac{v}{c}\cos\theta\right), \tag{2.34}$$

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 frictional force on the electron opposite its direction of motion. The CMB energy density within a solid angle $d\Omega = d\cos\theta\, d\phi$ (in spherical coordinates) will be $d\epsilon = a_{\rm rad}T^4(\theta)\, d\Omega/4\pi$ (where $a_{\rm rad}$ is the radiation constant). Since each photon carries a momentum equal to its energy divided by c, the electron will be slowed 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,

$$m_e \frac{dv}{dt} = -\int \sigma_T \cos\theta \, d\epsilon = -\frac{4}{3c}\sigma_T a_{\rm rad} T_\gamma^4 v. \tag{2.35}$$

The rate of energy loss by the electron is obtained by multiplying the drag force by v, which yields

$$\frac{d}{dt}E = -\frac{8\sigma_T}{3m_e c} a_{\rm rad} T_\gamma^4 E, \tag{2.36}$$

where $E = (1/2)\, m_e v^2$. For a thermal ensemble of electrons at a nonrelativistic temperature T, the average energy is $\langle E \rangle = (3/2)\, k_B T_e$. 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 that balances the 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 $\langle E \rangle = (3/2)\, k_B T_\gamma$, at which point the ensemble 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

$$\frac{dT_e}{dt} = \frac{x}{(1+x)}\left[\frac{T_\gamma - T_e}{t_{\rm C}(z)}\right] - 2HT_e, \tag{2.37}$$

where $t_{\rm C}$ is the Compton cooling time,

$$t_{\rm C} \equiv \left(\frac{8\sigma_T a_{\rm rad} T_\gamma^4}{3m_e c}\right)^{-1} = 1.2 \times 10^8 \left(\frac{1+z}{10}\right)^{-4} \ {\rm yr}, \tag{2.38}$$

and x is the fraction of all electrons that are free. For an electron–proton gas, $x = n_e/(n_e + n_{\rm H})$, where n_e and $n_{\rm H}$ are the electron and hydrogen densities respectively, and $T_\gamma \propto (1+z)$. The second term on the right-hand side of equation (2.37), $-2HT_e$, yields the adiabatic scaling $T_e \propto (1+z)^2$ in the absence of energy exchange with the CMB.

The relative importance of these two heating and cooling mechanisms therefore depends on the residual fraction of free electrons after cosmological recombination. If we ignore helium for simplicity, the rate at which electrons recombine is roughly[iii]

$$\frac{dx}{dt} = -\alpha_B(T_e) x^2 \bar{n}_{\rm H}, \tag{2.39}$$

[iii] At high redshifts, recombination is delayed by the large photon density and line emission. Detailed calculations at $z \gg 100$ require tracking the complex network of recombination reactions.

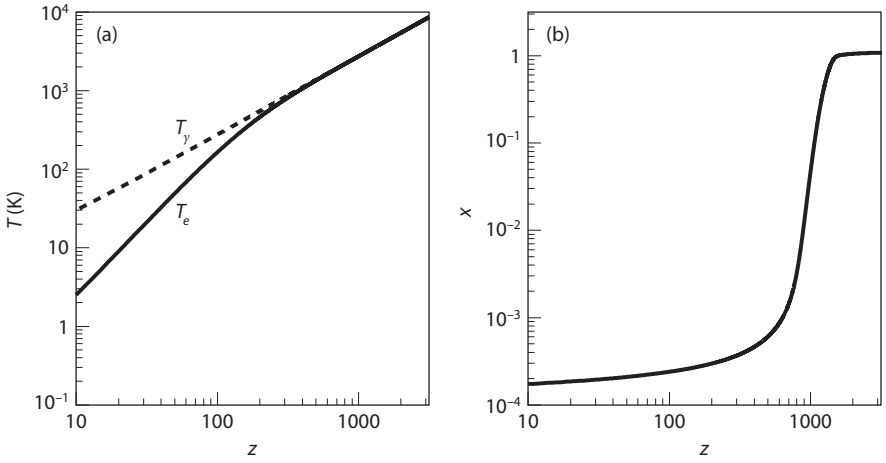


Figure 2.5 Thermal and ionization histories of the Universe before the first stars formed (panels a and b, respectively). In the top panel, the solid and dotted curves show T_e and T_γ, respectively. Note that the ionized fraction x decreases rapidly after recombination at $z \sim 1100$ and then "freezes out" at $z \sim 300$. Meanwhile, Compton scattering keeps $T_e \approx T_\gamma$ until $z \sim 200$, after which the declining CMB energy density and small residual ionized fraction are no longer sufficient to maintain thermal contact between the gas and CMB. At later times, $T_e \propto (1+z)^2$, as appropriate for an adiabatically expanding non-relativistic gas. These results were produced with the publicly available code RECFAST (http://www.astro.ubc.ca/people/scott/recfast.html).

left / dashed

where $\alpha_B \propto T_e^{-0.7}$ is the case-B recombination coefficient.[iv] With our preferred cosmological parameters, the fractional change in x per Hubble time is therefore

$$\frac{\dot{n}_e}{H n_e} \approx 7x(1+z)^{0.8}. \tag{2.40}$$

Electrons "freezeout" and cease to recombine effectively when this factor becomes on the order of unity; after that point, the Hubble expansion time is shorter than the recombination time. More precise numerical calculations give $x \approx 3 \times 10^{-4}$ at $z \approx 200$, as shown in Figure 2.5.

Inserting this value into equation (2.37), we find that the small fraction of residual electrons enforces thermal equilibrium between the gas and CMB down to $z \approx 200$, when Compton heating finally becomes inefficient. Figure 2.5 shows a more exact calculation: note how the gas and CMB temperatures begin to depart at $z \sim 200$, and the gas begins to follow the expected adiabatic evolution $T_e \propto (1+z)^2$ at $z \sim 100$.

[iv] This ignores recombinations to the ground state, which generate a new ionizing photon and so do not change the net ionized fraction. See §9.2.1 for more discussion of the recombination rate.

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

2.2.1 Fluctuations in the IGM Temperature

Equation (2.37) describes the evolution of the mean IGM temperature. However, two factors can induce inhomogeneities in this field. First, the CMB temperature varies slightly across the Universe, so each electron will scatter off a different T_γ. Second, the adiabatic expansion term depends on the local density. In an overdense region, where gravity slows the expansion (or even causes contraction), the cooling is slower (and may turn into heating); in an underdense region, the cooling accelerates. Thus, the IGM will be seeded by small temperature fluctuations that reflect its density structure.

To describe these fluctuations, we write δ_T as the fractional temperature fluctuation and δ_γ as the photon density fluctuation and note that (for a blackbody) $\delta_\gamma = 4\delta_{T_\gamma}$, where the latter is the photon temperature fluctuation. Then, the analog of equation (2.37) is

$$\frac{d\delta_T}{dt} = \frac{2}{3}\frac{d\delta_b}{dt} + \frac{x(t)}{t_C(z)}\left[\delta_\gamma\left(\frac{\bar{T}_\gamma}{\bar{T}_e} - 1\right) + \frac{\bar{T}_\gamma}{\bar{T}_e}(\delta_{T_\gamma} - \delta_T)\right]. \tag{2.41}$$

Here the first term describes adiabatic cooling due to expansion (allowing for variations in the expansion rate), and the second accounts for variations in the rate of energy exchange through Compton scattering (which can result from variations in either the gas or photon temperatures); overbars denote the mean values for the CMB and electron temperatures.

Meanwhile, the fluctuations in the baryon temperature influence the density evolution as well. If we allow arbitrary fluctuations in the temperature field, rather than forcing them to trace the density fluctuations, equation (2.6) then reads

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = \frac{3}{2}H^2(\Omega_c\delta_c + \Omega_b\delta_b) - \frac{k^2}{a^2}\frac{k_B\bar{T}_e}{\mu m_p}(\delta_b + \delta_T). \tag{2.42}$$

lc / PE

This, result together with equations (2.41), (2.37), and (2.39) for the temperature and ionized fraction evolution, provides a complete set of equations for tracing the density and temperature evolution, in the absence of relative streaming. If streaming is included, the final term in equation (2.27) must be replaced by the final term in equation (2.42).

Figure 2.6 shows the resulting power spectra for δ_c, δ_b, δ_T, and δ_{T_γ} at four different redshifts. Note how the photon perturbations are strongly suppressed on small scales (below the sound horizon) thanks to their large pressure. Near recombination, the baryonic perturbations are also suppressed on these scales, especially in the temperature. After recombination, the baryons fall into the dark matter potential wells, where their perturbations grow rapidly, and temperature fluctuations also grow quickly thanks largely to the variations in the

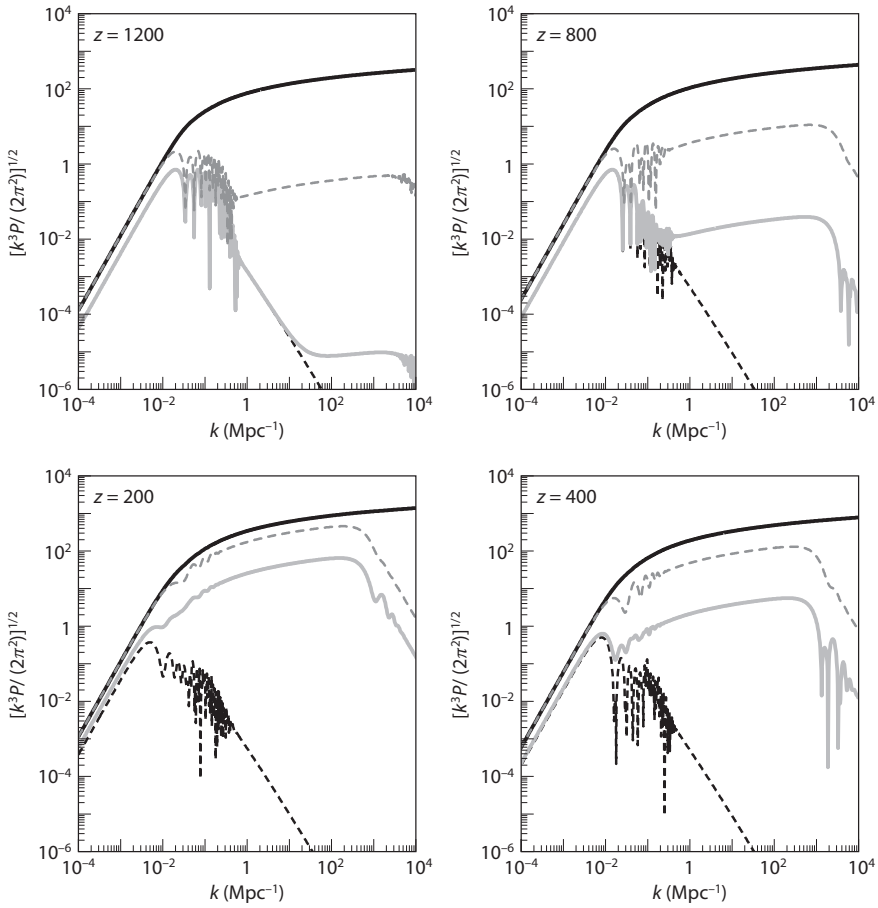


Figure 2.6 Power spectra for density and temperature fluctuations versus comoving wavenumber at four different redshifts. The curves show the CDM density (solid), baryon density (dotted), baryon temperature (short dashes) and photon temperature (long dashes). These curves do not include the relative streaming of the baryons and cold dark matter. Naoz, S., & Barkana, R., *Mon. Not. R. Astron. Soc.* **362**, 1047 (2005). Copyright 2005 by the Royal Astronomical Society.

adiabatic cooling rate. The turnover at very small scales in the baryonic power spectrum is due to the finite pressure of the gas. The baryon acoustic oscillations are also visible near $k \sim 0.01$ Mpc^{-1}.

AU: Since equation 2.11 doesn't have any corresponding equation, from equation 2.12 has been renumbered. Please check and confirm ok as set.

OK; requires renumbering equation references in later chapters (see many corrections in rest of text)

Chapter Three

Nonlinear Structure and Halo Formation

In the last chapter, we followed the evolution of structure in the linear regime, when the perturbations are small. Of course, most of the objects we study with telescopes are far outside this regime, with typical densities many thousands of times the cosmic mean. In this chapter, we take the next steps toward understanding these objects by studying the evolution of perturbations in the nonlinear regime. We focus for the most part on analytic models that shed light on the physical processes involved.

The advent of computer technology has made numerical studies of nonlinear evolution almost routine, and many of today's theoretical calculations follow this path. The analytic approaches we describe inform these calculations, but the numerical simulations allow us to sharpen our conclusions and predictions. We discuss this synergy and describe "semianalytic" models that can be written analytically but whose ultimate justification lies in their good agreement with numerical simulations. We describe the fundamental aspects of computational methods in the last section of the chapter.

3.1 Spherical Collapse

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

For simplicity, let us consider an isolated, spherically symmetric density or velocity perturbation of the smooth cosmological background and examine the dynamics of a test particle at a radius r relative to the center of symmetry. Birkhoff's theorem (see §1.2.2) implies that we may ignore the mass outside this radius in computing the motion of our particle. The equation of motion describing the system reduces to the usual Friedmann equation for the evolution of the scale factor of a homogeneous Universe, but with a density parameter Ω that now takes into account the additional mass interior to the shell and its modified expansion velocity. In particular, despite the arbitrary density and velocity profiles given to the perturbation, only the total mass interior to the particle's radius and the peculiar velocity at the particle's radius contribute to the effective value of Ω. We may thus find a solution to the particle's motion that describes

its departure from the background Hubble flow and its subsequent collapse or expansion. This solution holds until our particle crosses paths with one from a different radius, which happens rather late for most initial conditions.

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

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

where r is the radius in a fixed (not comoving) coordinate frame, H_0 is the present-day Hubble constant, and the unperturbed Hubble flow velocity (to which the previously mentioned peculiar velocity should be added) is given by $dr/dt = H(t)r$. The total mass enclosed within radius r is $M = (4\pi/3) r_i^3 \rho_i (1 + \delta_i)$, where ρ_i is the background density of the Universe at time t_i. We next define the dimensionless radius $x = a_i(r/r_i)$ and rewrite equation (3.1) as

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

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

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

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

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

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

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

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

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

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

where $a_{\max}$ is the value of a that sets the denominator of the integrand to zero, and we have used the fact that $\delta_i \ll 1$. (The integral itself determines the total expansion time; the factor of 2 accounts for the time from maximum expansion to collapse.) The analogy to a test particle escaping a point mass in equation (3.1) is illuminating here: in that case the constant K is simply proportional to the total energy per unit mass of the system, which determines whether the particle escapes to infinity. Here, a large negative K (enough to overcome the effective repulsive force from the cosmological constant) implies the same recollapse.

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. In fact, although we have cast this problem as a test particle in an overdense or underdense region, we could have developed exactly the same equations by carving out a spherical region from a truly uniform medium. Then, the three possibilities would simply correspond to a closed, an open, and a flat Universes (with $\Omega_\Lambda = 0$). The three solutions may be written as

needs space before (/ PE

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

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

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

where $A^3 = GMB^2$ applies in all cases even though the constants have different values in each one. All three solutions have $r^3 = 9GMt^2/2$ as t goes to zero, which matches the linear theory expectation that the perturbation amplitude get smaller as one goes back in time. In the closed case, the shell turns around at time πB and radius $2A$ (when its density contrast relative to the background of an $\Omega_m = 1$ Universe is $9\pi^2/16 = 5.6$), and collapses to zero radius at time $2\pi B$. Interestingly, these collapse times are independent of the initial distance from the origin: perturbations with fixed initial density contrast collapse homologously, with all shells turning around and collapsing at the same time. Figure 3.1 illustrates the stages of this collapse process.

This is the fully nonlinear solution for the simplified problem of collapse of a purely spherical top-hat perturbation. Of course, the real density distribution of the Universe is much more complicated. Although we cannot describe analytically the full nonlinear evolution of density perturbations, we *can* fully describe their linear evolution. A compromise is then to use this linear evolution to identify regions (such as galaxies) where spherical nonlinear evolution is not a bad

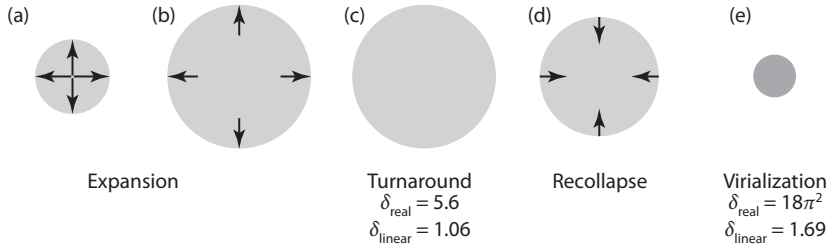


Figure 3.1 Stages of the spherical collapse model. At first, the overdensity (in gray) expands, though its excess gravity quickly slows that expansion below that of the Hubble flow. When the real fractional overdensity reaches $9\pi^2/16 \approx 5.6$ (corresponding to a *linearized* overdensity of 1.06), the expansion stops at *turnaround* and then begins to recollapse. When the overdensity reaches $18\pi^2 \approx 178$ (corresponding to a *linearized* overdensity of 1.69), the perturbation *virializes* as a collapsed dark matter halo.

approximation. It is therefore useful to determine the mapping between the *linear* density field described by perturbation theory and the *nonlinear* densities in the spherical model.

To do this, we are faced with the problem of relating the spherical collapse parameters A, B, and M to the linear theory density perturbation δ. This exercise is straightforward for the case of $\Omega_\Lambda = 0$ and $\Omega_m = 1$ or $K = 0$ ($K > 0$ [$K < 0$] produces an open [closed] model). By substituting equation (3.8) into equation (3.5) at the turnaround radius, we find

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

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

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

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

where $D(z)$ is the linear growth factor (see equation 2.7), although the true density (computed with the full nonlinear theory) differs. This critical density plays a key role in calculations of the halo abundance.

Of course, we do not expect a real object to collapse to a zero size; anisotropies and angular momentum in the initial distribution will prevent perfect collapse. Instead, we envision that the material will *virialize*, with strong particle interactions transforming the bulk kinetic energy of collapse into random velocities.

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

The result is a *dark matter halo* with a centrally concentrated mass distribution; we discuss the properties of such halos in §3.3

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

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

so that we may use $\overline{\delta_i}$ in place of δ_i in the preceding formulas. 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 reexpand, since these expanding shells will cross infalling shells. In such a case, first-time infalling shells will never be affected prior to their turnaround; the more complicated behavior after turnaround 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 at $z > 1$ when $\Omega_m \approx 1$.[1]

3.2 Cosmological Jeans Mass

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

We can obtain a rough estimate of the relative importance of these forces from the following simple considerations (see Figure 3.2). The increase in gas density near the center of a cloud sends out a pressure wave that propagates outward 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 timescale for the collapse of the cloud is given by its radius R divided by the free-fall speed $\sim(2GM/R)^{1/2}$, ~~whih~~ which / PE yields $t_{\rm coll} \sim (G\langle\rho\rangle)^{-1/2}$, where $\langle\rho\rangle = M/(4\pi R^3/3)$ ~~is the~~ characteristic density of the cloud as it turns around on its way to collapse.[ii]

If the sound wave does not have sufficient time to traverse the cloud during the free-fall time, namely $R > c_s t_{\rm coll}$, then the cloud will collapse. Under

[ii] Substitution of the mean density of the earth into this expression yields the characteristic time it takes a freely falling elevator to reach the center of the earth from its surface (~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).

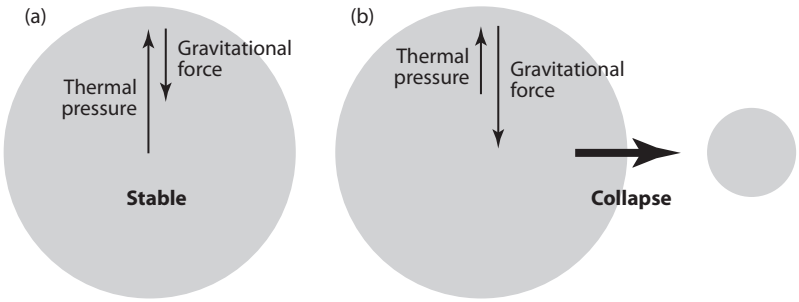


Figure 3.2 The Jeans criterion for gravitational collapse of baryonic material. In *(a)*, the outward pressure gradient exceeds the inward gravitational force, so the region is stable. In *(b)*, the opposite is true, and the resulting instability causes runaway collapse. Note that if the pressure changes over time, a time-averaged pressure restoring force must be used, as described in §3.2.1.

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. However, the collapse will be inhibited by pressure for a sufficiently small cloud with $R < c_s t_{\rm coll}$. The transition between these regimes is defined by the so-called Jeans radius, $R_{\rm J} \sim c_s t_{\rm coll}$, which determines how large a perturbation must be before gravitational instability triggers collapse.

More precisely, in a static, infinite, uniform gas (with density ρ and sound speed c_s) that obeys Newtonian gravity the *Jeans length* $\lambda_{\rm J}$ is defined as the critical wavelength that separates oscillatory and exponentially growing density perturbations. A detailed stability analysis yields

$$\lambda_{\rm J}^2 = \frac{\pi c_s^2}{G\rho}. \tag{3.15}$$

The *Jeans mass* is defined as the mass within a sphere of radius $\lambda_{\rm J}/2$, or

$$M_{\rm J} = \frac{4\pi}{3}\rho\left(\frac{\lambda_{\rm J}}{2}\right)^3. \tag{3.16}$$

This corresponds to associating the region in which a sine or cosine wave is positive (of width $\lambda_{\rm J}/2$) as the "object" that is collapsing. 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, the fundamental concept of a minimum mass that can collapse at any given time remains the same.

We follow the notation of §2.1 and consider a mixture of dark matter and baryons with density parameters $\Omega_c(z)$ and $\Omega_b(z)$, respectively. We write $\Omega_m(z)$ for the total matter density. We will examine the evolution of a single Fourier mode on a scale much smaller than the horizon. In comoving coordinates, we then want to solve equation (2.6) for the critical wavenumber for collapse. (Here for simplicity we are ignoring the relative streaming of the baryons and dark matter as well as temperature fluctuations, but see further.) The Jeans wavelength $\lambda_{\rm J} = 2\pi/k_{\rm J}$ is obtained by setting the right hand side of equation (2.6) to zero, and solving for the critical wavenumber $k_{\rm J}$:

$$k_{\rm J} = \frac{2}{3}\frac{aH}{c_s}. \tag{3.17}$$

The critical wavelength $\lambda_{\rm J}$ (and therefore the mass $M_{\rm J}$) is clearly time dependent in general.

For a concrete estimate, we need an expression for the sound speed. We adopt an ideal gas equation of state for the baryons with an adiabatic index (or specific heat ratio) $\gamma = 5/3$. Initially, at time $t = t_{\rm i}$, we suppose that the gas temperature is uniform $T_{\rm b}(R, t_{\rm i}) = T_{\rm i}$. As described in §2.2, at very high redshifts the baryon temperature traces the CMB temperature, $T_{\rm b} \propto T_\gamma \propto (1+z)$, while at $z < z_t \sim 100$ baryons cool adiabatically, $T_{\rm b} \propto \rho_b^{(\gamma-1)} \propto (1+z)^2$. We will therefore account for both possibilities with a parameter β_T so that $T_{\rm b} \propto (1+z)^{\beta_T - 1}$. In that case, we can include gas temperature perturbations as well and write equation (2.43) as

$$\ddot{\delta}_b + 2H\dot{\delta}_b = \frac{3}{2}H^2\left(\Omega_b\delta_b + \Omega_c\delta_c\right) - \frac{kT_{\rm i}}{\mu m_p}\left(\frac{k}{a}\right)^2\left(\frac{a_{\rm i}}{a}\right)^{(1+\beta_T)} \times \left(\delta_b + \frac{2}{3}\beta_T[\delta_b - \delta_{b,i}]\right). \tag{3.18}$$

Here $\mu = 1.22$ is the mean atomic weight of the neutral primordial gas in units of the proton mass. The last term on the right-hand side 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 that develops in the adiabatic limit. We infer from equation (3.18) that as time proceeds, perturbations with increasingly smaller initial wavelengths stop oscillating and start to grow.

To estimate this scale, we assume $\delta_b \sim \delta_c$ and consider sufficiently high redshifts at which the Universe is matter dominated (so that $\Omega_m \approx 1$). Following cosmological recombination at $z \approx 10^3$, the residual ionization of the cosmic gas keeps its temperature locked to the CMB temperature (via Compton scattering) down to a redshift of $z \approx 200$ (see §2.2 and Figure 2.5) In this redshift range, $\beta_T = 0$ and

$$k_{\rm J} \equiv (2\pi/\lambda_{\rm J}) = [2kT_{\gamma,0}/3\mu m_p]^{-1/2}\sqrt{\Omega_m}H_0, \tag{3.19}$$

insert line break in equation here, before minus sign / PE

insert subscript italics capital B / AA (x2)

where $T_{\gamma,0}$ is the CMB temperature today. Then, the Jeans mass is redshift independent and obtains a value (for the total mass of baryons and dark matter)

$$M_{\rm J} \equiv \frac{4\pi}{3}\left(\frac{\lambda_{\rm J}}{2}\right)^3 \bar{\rho}_c = 1.35\times 10^5 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} M_\odot, \tag{3.20}$$

where $\bar{\rho}_c$ is the average comoving matter density.

At $z < 100$, the gas temperature declines adiabatically with $\beta_T = 1$, and the total Jeans mass obtains the value

$$M_{\rm J} = 4.54\times 10^3 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} \left(\frac{\Omega_b h^2}{0.022}\right)^{-3/5} \left(\frac{1+z}{10}\right)^{3/2} M_\odot. \tag{3.21}$$

These equations set the minimum threshold mass for gas clumps to have formed in the early Universe. We emphasize that the values depend on the IGM temperature, which we have assumed either traces the CMB or cools adiabatically. We will see in later chapters that radiative processes—especially photoionization—can strongly influence the IGM temperature, increasing the Jeans mass dramatically.

So far, we have ignored similar effects in the dark matter component: although these collisionless particles do not feel a pressure force, their intrinsic velocity dispersion plays a role analogous to pressure, and a similar criterion for collapse exists. However, in popular *cold* dark matter models with weakly interacting massive particles, the Jeans mass of the dark matter alone is negligible but nonzero, of the order of the mass of a planet like Earth or Jupiter.[2] All halos between this minimum clump mass and $\sim 10^5\, M_\odot$ are expected to contain mostly dark matter and little ordinary matter. *Warm* dark matter, with a moderately large velocity dispersion, could change this expectation and—if its Jeans mass exceeds that of the baryons—substantially modify the early phases of structure formation.

3.2.1 The Filtering Mass

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

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

$$\delta_b(t,k) = \frac{\delta_c(t,k)}{1+k^2/k_{\rm J}^2}, \tag{3.22}$$

where the dark matter fluctuation grows in proportion to the linear growth factor, $\delta_c \propto D(t)$ (equation 2.7). In the general case where the Jeans wavenumber

k_J is time dependent, we can identify the proper time averaging by considering the perturbative effect of gas pressure on large scales. We therefore expand the ratio $\delta_b(t,k)/\delta_c(t,k)$ in powers of k^2 with $\delta_b(t,k=0) = \delta_c(t,k=0)$. The ratio between the linear overdensity of the baryons and dark matter in the limit of small k can then be written as

$$\frac{\delta_b}{\delta_c} = 1 - \frac{k^2}{k_F^2} + ..., \tag{3.23}$$

insert comma / PE

or, equivalently as

$$\frac{\delta_b(t,k)}{\delta_c(t,k)} = 1 - \frac{A(t)}{D(t)}k^2, \tag{3.24}$$

where $A(t) \equiv D(t)/k_F^2$ can be found by substituting the latter relation into the coupled linear growth equations for δ_b and δ_c and ignoring terms of order k^4 or higher (equation 2.6). This substitution gives the differential equation

$$\frac{d^2A}{dt^2} + 2H\frac{dA}{dt} = \frac{c_s^2}{a^2}D(t). \tag{3.25}$$

The filtering wavenumber k_F is the solution to this equation. Writing it in terms of the Jeans wavenumber k_J (using the latter's relation to c_s in equation 3.17), we obtain

$$\frac{1}{k_F^2(t)} = \frac{1}{D(t)}\int_0^t dt'\, a^2(t')\frac{\ddot{D}(t') + 2H(t')\dot{D}(t')}{k_J^2(t')}\int_{t'}^t \frac{dt''}{a^2(t'')}\,. \tag{3.26}$$

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

$$\frac{1}{k_F^2(t)} = \frac{3}{a}\int_0^a \frac{da'}{k_J^2(a')}\left(1 - \sqrt{\frac{a'}{a}}\right)\,. \tag{3.27}$$

Figure 3.3 contrasts the time-averaged filtering mass (computed using a full perturbative analysis, without the simplifications of equation 3.27) with the instantaneous Jeans mass (thin solid and dot-dashed curves, respectively). Note how the Jeans mass declines with cosmic time as the Universe cools, but the filtering mass remains roughly constant over this wide redshift interval. The filtering mass is initially much smaller than the Jeans mass because the baryon fluctuations are suppressed after recombination and must then catch up to the dark matter. But the filtering mass remains roughly constant with redshift, whereas the Jeans mass declines rapidly as the Universe expands and cools. By $z \sim 10$, the Jeans mass is several times smaller than the filtering mass.

It is conventional to assume that the Jeans or filtering mass accurately reflects the threshold for baryonic structure formation. However, linear theory specifies only whether an initial perturbation, characterized by the parameters k, $\delta_{Cm,i}$, δ_b, i, and t_i, begins to grow, and the preceding analysis is perturbative (valid only as long as δ_b and δ_c are much smaller than unity). As δ_b and δ_c grow and become larger than unity, the density profiles start to evolve, and dark matter shells may cross baryonic shells owing to their different dynamics. Hence the amount of mass enclosed within a given baryonic shell may increase with

should be greek lowercase delta subscript italics c roman ,i / PE

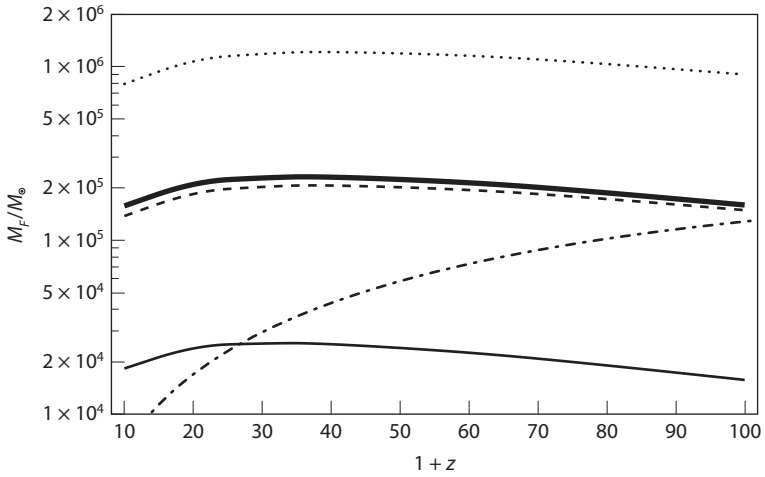


Figure 3.3 The Jeans and filtering masses as a function of redshift. The dot-dashed curve shows the Jeans mass computed with equation (3.17), while the thin solid curve shows the filtering mass from equation (3.27). The other lines include baryon streaming: the dashed and dotted take $v_{\rm bc} = \langle v_{\rm bc}^2 \rangle^{1/2}$ and $2\langle v_{\rm bc}^2 \rangle^{1/2}$, while the thick solid line averages over the entire velocity distribution. Tseliakhovich, D., Barkana, R., & Hirata, C., *Mon. Not. R. Astron. Soc.* **418**, 906 (2011). Copyright 2011 by the Royal Astronomical Society.

time, until eventually the dark matter pulls the baryons with it and causes their collapse even for objects nominally below the filtering mass. To determine the minimum mass of the resulting nonlinear baryonic object following the shell crossing and virialization of the dark matter, we typically appeal to the spherical collapse model described in the previous section.

3.2.2 Primordial Streaming of Baryons and Halo Formation

As we saw in §2.1.2, the baryonic gas also has a substantial velocity relative to the baryons at early times. Figure 2.2 shows the variance of the velocity difference perturbations (in units of c) per $\ln k$ as a function of the mode wavenumber k at $z = 10^3$. The power declines rapidly at $k > 0.5$ Mpc^{-1}, indicating that the velocity of the baryons relative to the dark matter was coherent over the photon diffusion (Silk damping) scale of several megaparsecs. This scale is much larger than the filtering mass, so to these halos the background intergalactic baryons appeared to be moving coherently as a wind. It is therefore important to examine whether this wind had a significant effect on the assembly of baryons onto the earliest galaxies.

By $z \sim 50$, the typical streaming velocity $\sim$1.5 km s^{-1} corresponded to an equivalent gas temperature $T_{\rm bc} \sim m_p v_{\rm bc}^2/k_B = 270\ {\rm K}[(1+z)/50]^2$, well above the IGM temperature. Effectively, the relative velocity acts as an increased sound speed (since it needs to be dissipated on virialization of the gas) and

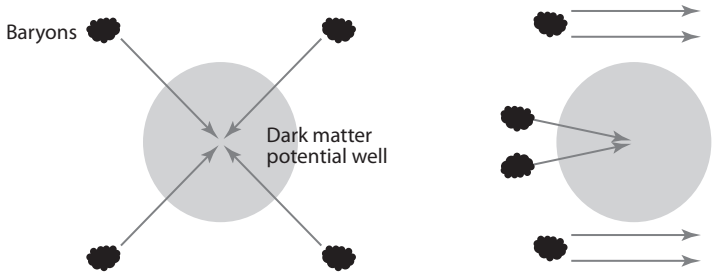


Figure 3.4 Cartoon representation of the effects of the relative streaming of baryons and dark matter on halo formation. In the absence of these motions (*left*), dark matter potential wells (which formed before recombination) attract nearby baryons, quickly building up their gas content so long as their total mass exceeds the filter mass. But when these motions are present (*right*), the baryons may move so quickly that they do not respond strongly to the dark matter potential unless they happen to pass very close to its center. In this case, baryonic accretion onto dark matter halos can be strongly suppressed (and can even affect dark matter accretion, by decreasing the total halo mass).

prevents the baryons from settling into the shallowest potential wells: very roughly, the sound speed in equations (2.6) and (3.17) should be replaced by $v_{\rm eff}^2 \equiv (v_{\rm bc}^2 + c_s^2)$ to account for this. Figure 3.4 illustrates this correction in a cartoon fashion.

The effect on the filtering mass is quite dramatic. Following the perturbative procedure outlined in §3.2.1 with the full set of linear perturbation equations (~~2.25~~)–(~~2.28~~), and including the temperature fluctuations as in §2.2.1, we can then determine the effective filtering mass as a function of the streaming velocity $v_{\rm bc}$. The thick solid curve in Figure 3.3 shows this quantity averaged over the full streaming velocity distribution (the dashed and dotted lines show the results at specific velocities). These increase the minimum mass for baryonic accretion by nearly an order of magnitude (on average) over naive expectations, though of course the actual value will depend on the local streaming velocity. (We will return to this suppression in the next section from a slightly different angle; see Fig. 3.9.)

2.24 / 2.27 / PE (from renumbering equations in ch. 2)

3.3 Halo Properties

When an object above the Jeans mass collapses, the dark matter forms a halo inside of which the gas may cool, condense to the center, and eventually fragment into stars. The dark matter cannot cool, since it has very weak interactions. As a result, a galaxy emerges with a central core occupied by stars and cold gas and is surrounded by an extended halo of invisible dark matter. Because 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 (∼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 that is *an order of magnitude smaller* than the size of the dark matter halo (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 (see §8.4), during which it tends to break into dense clouds that 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 the order of the mass of a star. As a result, the clumps fragment into stars, and a galaxy is born; we discuss the physics of this process in Part II.

These dark matter halos therefore form the fundamental units within which astronomical objects form, and it is important to understand their properties. By solving the equation of motion (3.1) for a spherical overdense region, we can relate the characteristic radius and gravitational potential well of each of these galaxies to their mass and their redshift of formation. In principle, a spherical region would collapse to a point mass, but of course the real world is not so idealized. As already mentioned, even a slight violation of the exact symmetry of the initial perturbation can prevent the top hat from collapsing entirely. Instead, the halo reaches a state of virial equilibrium through violent dynamical relaxation. We are familiar with the fact that the circular orbit the Earth around the Sun has a kinetic energy that 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). Within the context of the spherical collapse model in §3.1, let us suppose that virialization occurs when the spherical perturbation would otherwise collapse completely, at twice the turnaround time. In a flat matter-dominated universe (appropriate at the high redshifts of most interest to us), $a \propto t^{2/3}$, so the Universe would have expanded by a factor $2^{2/3}$ between turnaround and virialization, and its mean density would have fallen by a factor of 4. Thus the final density of the virialized object relative to the critical density is

$$\Delta_{\rm vir}(\Omega_m = 1) \equiv \frac{\rho_{\rm vir}(z_{\rm vir})}{\bar{\rho}_{\rm crit}(z_{\rm vir})} = \left(\frac{9\pi^2}{16}\right) \times 8 \times 4 = 18\pi^2 \approx 178, \qquad (3.28)$$

where the first factor is the overdensity at turnaround (see §3.1), the second is the inverse of the change in volume from turnaround, and the last is the change in the mean density of the Universe over the collapse interval. Note that the virial overdensity at collapse implies that the dynamical time within the virial radius of galaxies, $\sim (G\rho_{\rm vir})^{-1/2}$, is on the order of a tenth of the age of the Universe at any redshift.

In a Universe with $\Omega_m + \Omega_\Lambda = 1$ the virial overdensity at the collapse redshift has the fitting formula[4]

$$\Delta_c = 18\pi^2 + 82d - 39d^2\,, \qquad (3.29)$$

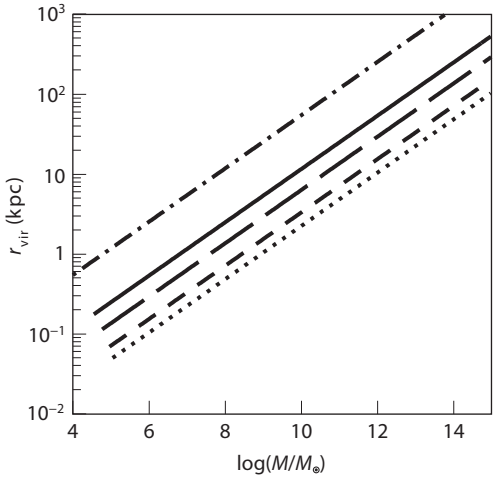


Figure 3.5 Virial radius over a wide range of halo masses and redshifts (equation 3.30). The dotted, short-dashed, long-dashed, solid, and dot-dashed curves take $z = 30,\ 20,\ 10,\ 5$, and 0, respectively.

where $d \equiv \Omega_m(z) - 1$ is evaluated at the collapse redshift, and $\Omega_m(z)$ is defined in equation (2.9).

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

$$r_{\rm vir} = 0.784 \left[\frac{\Omega_m}{\Omega_m(z)}\,\frac{\Delta_c}{18\pi^2}\right]^{-1/3} \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{-1} h^{-1/2}\ {\rm kpc} \qquad (3.30)$$

-2/3 /AA

and a corresponding circular velocity

$$V_c = \left(\frac{GM}{r_{\rm vir}}\right)^{1/2} = 23.4 \left[\frac{\Omega_m}{\Omega_m(z)}\,\frac{\Delta_c}{18\pi^2}\right]^{1/6} \left(\frac{M}{10^8 M_\odot}\right)^{1/3} \times \left(\frac{1+z}{10}\right)^{1/2} h^{-1/3}\ {\rm km\ s^{-1}}. \qquad (3.31)$$

eliminate `-' should be 1/3 /AA

Please line "x" up with 23.4 on above line / PE

We may also define a virial temperature

$$T_{\rm vir} = \frac{\mu m_p V_c^2}{2k} = 1.98 \times 10^4 \left(\frac{\mu}{0.6}\right) \left[\frac{\Omega_m}{\Omega_m(z)}\,\frac{\Delta_c}{18\pi^2}\right]^{1/3} \times \left(\frac{M}{10^8 M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) h^{2/3}\ {\rm K} \qquad (3.32)$$

insert subscript italics capital B / AA

eliminate `-' should be 2/3 /AA

Note that the value of μ depends on the ionization fraction of the gas; a fully ionized primordial gas has $\mu = 0.59$, while a gas with ionized hydrogen but only singly ionized helium has $\mu = 0.61$, and a fully neutral primordial gas has $\mu = 1.22$. For context, Figures 3.5, 3.6, and 3.7 show these quantities over a wide range of halo masses and redshifts.

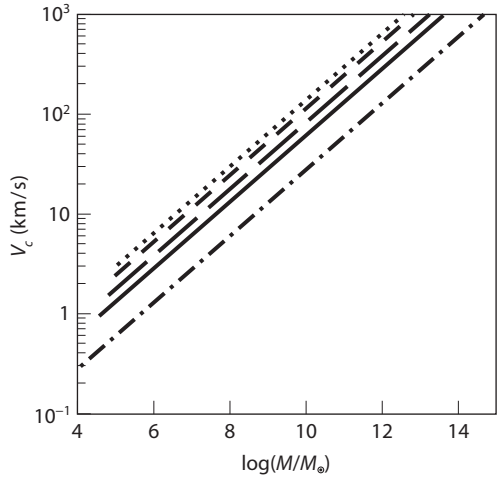


Figure 3.6 Circular velocity over a wide range of halo masses and redshifts (equation 3.31). The dotted, short-dashed, long-dashed, solid, and dot-dashed curves take $z = 30,\ 20,\ 10,\ 5$, and 0, respectively.

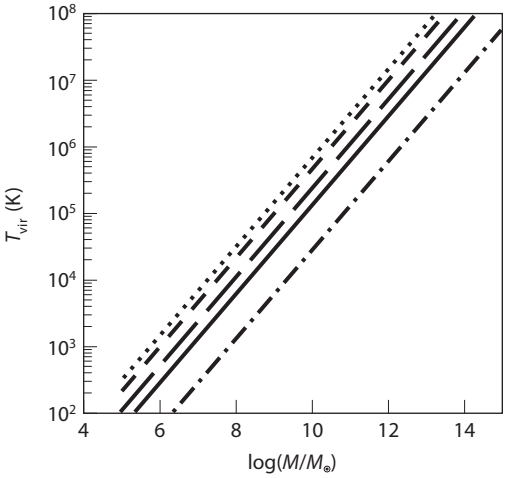


Figure 3.7 Virial temperature over a wide range of halo masses and redshifts (equation 3.30). The dotted, short-dashed, long-dashed, solid, and dot-dashed curves take $z = 30,\ 20,\ 10,\ 5$, and 0, respectively. For simplicity we assume fully ionized gas in all these halos, though the low-mass objects may actually accrete neutral gas.

The binding energy of the halo is approximately

$$E_b = \frac{1}{2}\frac{GM^2}{r_{\rm vir}} = 5.45 \times 10^{53} \left[\frac{\Omega_m}{\Omega_m(z)}\,\frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{M}{10^8 M_\odot}\right)^{5/3} \left(\frac{1+z}{10}\right) h^{-2/3}\,{\rm erg} \quad (3.33)$$

eliminate `-' should be 2/3 /AA

If necessary to break line, please insert "x" and line up with 5.45

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 ($\sim 10^{51}$ erg) for halos near the filtering mass (although, as we will see later, cooling is likely to increase this binding energy by quite a large factor).

Although spherical collapse captures some of the physics governing the formation of halos, structure formation in cold dark matter models proceeds hierarchically. At early times, most of the dark matter was in low-mass halos, and these halos then continuously accreted and merged to form high-mass halos. Numerical simulations of hierarchical halo formation indicate a roughly universal spherically averaged density profile for the resulting halos, though with considerable scatter among different halos. This so-called NFW (or Navarro-Frenk-White) profile has the approximate form[5]

$$\rho(r) = \frac{3H_0^2}{8\pi G}(1+z)^3 \frac{\Omega_m}{\Omega_m(z)} \frac{\delta_c}{c_{\rm N} x (1+c_{\rm N} x)^2}\ , \tag{3.34}$$

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

$$\delta_c = \frac{\Delta_c}{3} \frac{c_{\rm N}^3}{\ln(1+c_{\rm N}) - c_{\rm N}/(1+c_{\rm N})}. \tag{3.35}$$

The concentration parameter itself depends weakly on the halo mass M and more strongly on the formation redshift, with a value ~ 4 for newly collapsed halos and a larger value < 20 at later times.

An even better (but more complex) fit to state-of-the-art CDM simulations is obtained with the Einasto profile[6],

$$\ln\left[\frac{\rho(r)}{\rho_{-2}}\right] = -\frac{2}{\alpha}\left[\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right], \tag{3.36}$$

where $\alpha \approx 0.16$ and r_{-2} is the radius where the logarithmic slope of the density profile equals the isothermal sphere value, $(d\ln\rho/d\ln r) = -2$. At this radius $r^2\rho$ peaks at a density value of $\rho_{-2} = \rho(r_{-2})$. For the NFW profile, $r_{-2} = r_{\rm vir}/c_{\rm N}$. We show these two profiles in Figure 3.8; note that they differ substantially only close to the halo center.

As an example of the utility of these parameterizations, consider the effects of the relative streaming of baryons and dark matter on gas accretion into halos (as in §3.2.2). As described earlier, the effective pressure resisting collapse is parameterized by the mass density times $v_{\rm eff}^2 = c_s^2 + v_{\rm bc}^2$. Gravity will overcome this effective pressure only if the crossing time $r/v_{\rm eff}$ exceeds the free-fall time, $1/\sqrt{G\rho_{\rm vir}} \sim r/V_c$, or, equivalently, if $v_{\rm eff} < V_c$ for the halo. Figure 3.9 shows a similar estimate for the minimum halo mass required to accrete baryonic material in the presence of streaming. The streaming velocity marked on the vertical axis is evaluated at $z = 100$, with $v_{\rm bc} \propto (1+z)^{-1}$ as the gas falls into the dark matter potential wells at lower redshifts. The horizontal line marks the expected $\langle v_{\rm bc}^2 \rangle^{1/2} = 3$ km s^{-1} at $z = 100$. Each curve shows the minimum mass at the labeled redshift for a reasonable range of streaming

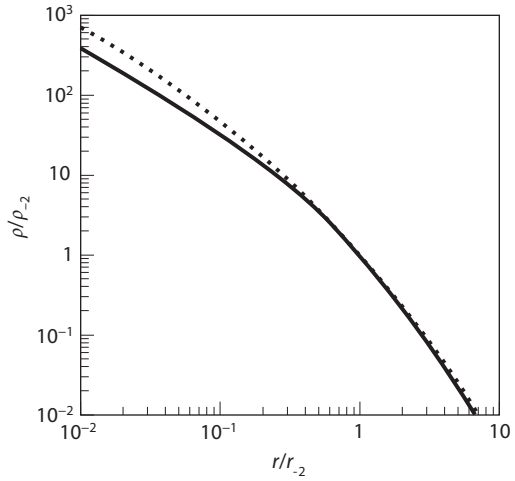


Figure 3.8 Parameterizations for the "universal" dark matter halo profile. The NFW prescription (equation 3.34) is shown with the solid line, while the Einasto profile (equation 3.36) is shown with the dotted line. We scale both to the point at which the logarithmic slope equals -2. For the NFW profile, this is at $r = r_{\rm vir}/c_{\rm N}$, where $c_{\rm N}$ is the concentration.

velocities; the symbols show the results of detailed numerical simulations at $z = 24$ (squares) and $z = 14$ (diamonds), which match this simple estimate reasonably well. Note that the streaming has little effect on halos with $M > 10^6\ M_\odot$ and is more important at higher redshifts, where the streaming velocities are larger.

3.4 Abundance of Dark Matter Halos

PE

In addition to characterizing the properties of individual halos, it is critical that any theory of structure formation predict 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 that successfully matches most of the numerical simulations was developed by Bill Press and Paul Schechter in 1974.[7] The model is based on the ideas of a Gaussian random field of density perturbations, linear gravitational growth, and spherical collapse. Once a region on the mass scale of interest reaches the threshold amplitude for collapse according to linear theory, it can be declared a virialized object. Counting the number of

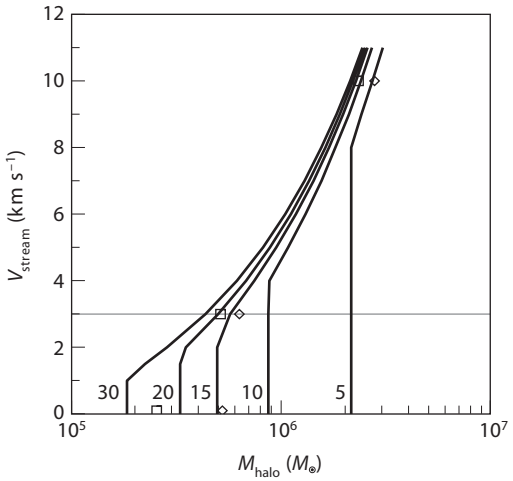


Figure 3.9 Effect of baryon–dark matter streaming on the minimum (Jeans) mass of a halo into which gas can assemble later and form stars (with the horizontal bar at 3 km s^{-1} marking the expected rms value). Each line represents the necessary halo mass for baryon collapse at the labeled redshift. Diamonds represent the final halo masses found in collapse simulations of a region with $\sigma_8 = 0.9$ ($z = 14$ with no streaming), squares represent masses from accelerated collapse simulations of a region with $\sigma_8 = 1.4$ ($z = 24$ with no streaming), and the lines delineate the prediction of the simple analytic model described in the text. The halo masses do not increase significantly at low streaming velocities. Halos collapsing at high redshift are more affected by relative streaming, as the physical streaming velocities are higher at these early times. Stacy, A., Bromm, V., & Loeb, A., *Astrophys. J.* **730**, L1 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

such density peaks per unit volume is straightforward for a Gaussian probability distribution.

To determine the abundance of halos at a redshift z, we use δ_M, the density field smoothed on a mass scale M, as defined in §2.1.1. Since δ_M is distributed as a Gaussian variable with zero mean and a standard deviation $\sigma(M)$ (which depends only on the linear power spectrum; see equation 2.34), the probability that δ_M is greater than some fixed δ equals

2.32

$$\int_\delta^\infty d\delta_M \frac{1}{\sqrt{2\pi}\,\sigma(M)} \exp\left[-\frac{\delta_M^2}{2\,\sigma^2(M)}\right] = \frac{1}{2}\mathrm{erfc}\left(\frac{\delta}{\sqrt{2}\,\sigma(M)}\right) . \qquad (3.37)$$

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

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

The second key ingredient is to note that even regions with $\delta_M < 0$ can actually be part of collapsed objects, if they are part of a region with $\delta > \delta_{\rm crit}$ on a scale $M' > M$. The original Press and Schechter paper solved this in an ad hoc fashion by multiplying the fraction in equation (3.37) by a factor of; 2 this guarantees that every dark matter particle is part of a halo (of some $M > 0$) even if its immediate environment is underdense. Thus, the final formula for the mass fraction in halos above M at redshift z, or the *collapse fraction* is

$$f_{\rm coll}(> M|z) = {\rm erfc}\left(\frac{\delta_{\rm crit}(z)}{\sqrt{2}\,\sigma(M)}\right). \tag{3.38}$$

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

One should note that at a redshift z, the probability distribution of δ_M is actually a Gaussian with variance $\sigma^2(M)\,D^2(z)$, because σ^2 is conventionally evaluated at the present day. The growth factor $D^2(z)$ then scales this to the actual density field at redshift z, provided we use the usual normalization $D(z=0) = 1$. This is possible because the growth factor is independent of physical scale: all Fourier modes grow by an identical factor during the matter era according to linear theory. However, we conventionally associate these redshift-dependent factors with $\delta_{\rm crit}$. Essentially, we imagine that we are working entirely with the linearized density field at $z = 0$ and allowing the threshold over which halos form to vary with redshift.

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

$$n(M) = \sqrt{\frac{2}{\pi}}\,\frac{\rho_m}{M}\,\frac{-d(\ln\sigma)}{dM}\,\nu_c\,e^{-\nu_c^2/2}\,, \tag{3.39}$$

where $\nu_c = \delta_{\rm crit}(z)/\sigma(M)$ is the number of standard deviations away from zero that the critical collapse overdensity represents on mass scale M and ρm is the comoving matter density. Note that $n(M)$ is often written as dn/dM in order to emphasize that it is a differential mass function, but we will generally use the more compact notation in this book.

Thus, the halo abundance depends on the two functions $\sigma(M)$ and $\delta_{\rm crit}(z)$, each of which depends on the fundamental cosmological parameters and the initial conditions of inflation. Figure 3.10 shows some example halo mass distributions at a variety of redshifts; note that the figure uses a slightly improved version of the mass function that we will discuss in the next sections.

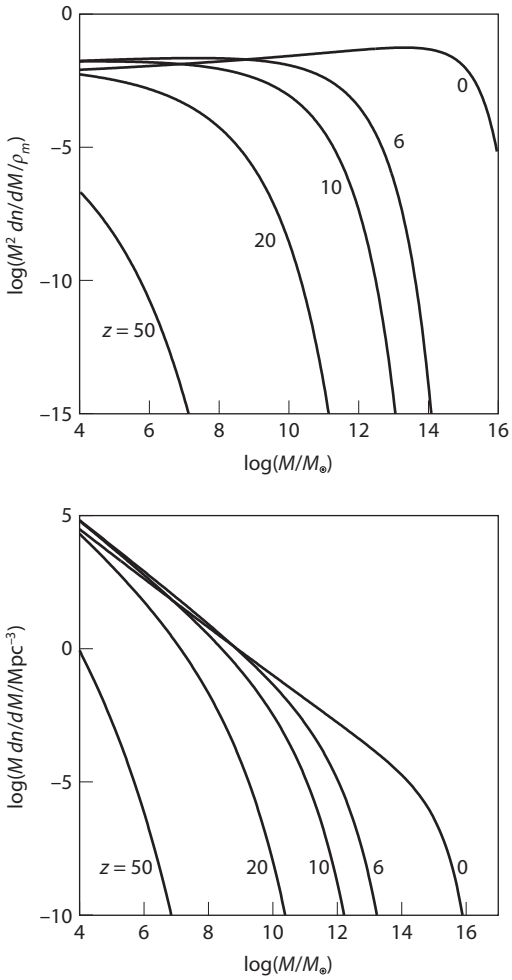


Figure 3.10 *Top:* The mass fraction incorporated into halos per logarithmic bin of halo mass, $(M^2 dn/dM)/\rho_m$, as a function of M at different redshifts z. Here $\rho_m = \Omega_m \rho_c$ is the present-day matter density, and $n(M)dM$ is the comoving density of halos with masses between M and $M+dM$. The halo mass distribution was calculated based on an improved version of the Press-Schechter formalism for ellipsoidal collapse that better fits numerical simulations (see §3.4.3). *Bottom:* comoving number density of halos per logarithmic bin of halo mass, Mdn/dM (in units of Mpc^{-3}), at various redshifts.

Insert comma / AA

3.4.1 The Excursion Set Formalism

Although the original Press-Schechter model is founded on an important physical insight, it turns out to be profitable to rephrase the problem in an entirely different way.[8] This provides two benefits: first, it yields a much more satisfying

derivation of the factor of 2 correction, and second, it provides a number of new insights into the spatial distribution and histories of dark matter halos.

among / PE

In particular, the Press-Schechter formalism makes no attempt to describe the correlations halos or between different mass scales. For example, 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. For many applications, we 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 would have to utilize the full spherically averaged density profile around a particular point. The potential correlations between the mean overdensities at different radii make the statistical description substantially more difficult.

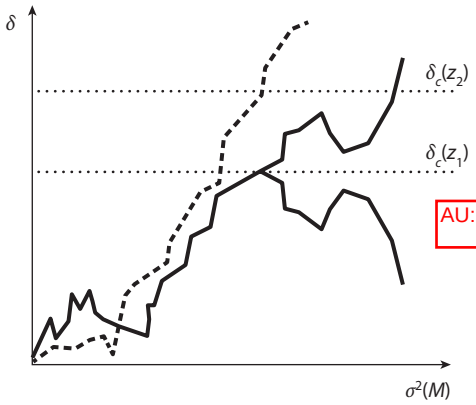
The excursion set formalism seeks to describe the statistics of halos by considering the statistical properties of δ_M, the average overdensity within some spherical window of characteristic mass M, as a function of M (or, equivalently, R). While the Press-Schechter model depends only on the Gaussian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

approach / PE

The basic idea is to view the density field around a given point, smoothed on different scales, as a diffusion process. Smoothed over a sufficiently large mass, $\delta_M \to 0$. As we zoom in to smaller scales, we naturally expect δ_M to deviate from zero, with a variance that must equal $\sigma^2(M)$. It is easiest to view this process in Fourier space: as we approach smaller scales, more and more Fourier modes become important, adding fluctuations to the density field. The particular set of modes contributing to a given spatial point will determine the "trajectory" of δ_M as a function of smoothed mass. The key insight of the excursion set approach is that we can consider this trajectory as a diffusion process (because each k-mode is independent of all others) and thereby calculate its statistics. Conceptually, each set of Fourier modes that one adds as M decreases provides a step in the random walk of the density field, so the key to a quantitative understanding of halo abundances is in generating the distribution of these random walks. Figure 3.11 illustrates this sequence for a few different sample points (or trajectories). The point will be part of a halo of mass M if its trajectory in this random walk first passes above the virialization threshold at that mass M.

2.31 / from renumbering ch. 2

The subtlety in this approach lies in defining the smoothed density field; recall that it is the full (linearized) density field convolved with a window function $W(R)$ (see equation 2.32). For most choices of window function, the quantities δ_M are correlated from one M to another so that it is difficult to calculate the desired statistics directly. However, for one particular choice of a window function, the correlations between different M greatly simplify, and many interesting quantities may be calculated.[9] We take advantage of the fact that in linear theory, each Fourier mode evolves independently, with no correlations between different scales k, and we use a k-space top-hat window function,



AU: Art ok before callout?

Ed: Yes, OK

Figure 3.11 Several random walk trajectories for the overdensity as a function of $\sigma^2(M)$, assuming a sharp k-space filter (or uncorrelated step sizes). The scale is arbitrary here; δ_c denotes a threshold of interest (i.e., $\delta_{\rm crit}$ in the Press-Schechter model). Note the wide variation between the trajectories and the rapid variations with scale: a particular trajectory may cross the threshold on multiple scales. In the excursion set model, however, we consider the barrier to be *absorbing*, so that a trajectory is marked by its first up-crossing. The two solid trajectories have equal probabilities of occuring which allows us to compute the halo abundance using the "mirror" trick described in the text.

insert comma / PE

namely, W_k = constant for all k less than some critical k_c and $W_k = 0$ for $k > k_c$. In that case, each step in the random walk corresponds to increasing k_c. For this filter,

$$\delta_M \propto \int_{k<k_c(M)} \frac{d^3k}{(2\pi)^3}\delta_k, \tag{3.40}$$

which means that the overdensity on a particular scale is simply the sum of the random variables (each Gaussian distributed) δ_k interior to the chosen k_c.

Because the filter is sharp, we simply add new Fourier modes to change the scale M. In linear theory, these evolve independently of the larger-wavelength modes already inside the filter, so the difference between δ_{M_1} and δ_{M_2} (where $M_1 > M_2$) is statistically independent of the value on the larger mass scale: i.e., each "step" in the walk is uncorrelated with previous steps. In particular, the difference is just the sum of the δ_k in the spherical k-shell between $k_c(M_1)$ and $k_c(M_2)$, which is independent of the sum of the δ_k interior to the smaller $k_c(M_1)$. We thus have a simple random walk, albeit one where the step size varies with k_c. Meanwhile, the distribution of δ_M (given no prior information about the random walk at larger M) is still a Gaussian of zero mean and a variance of $\sigma^2(M)$ (see equation 2.33).

2.32 / from renumbering ch. 2

Unfortunately, this filter is fundamentally inconsistent with the threshold $\delta_{\rm crit}$. The k-space top-hat filter has a *spatial* form $W(r) \propto j_1(k_c r)/k_c r$, where

2.32 / from renumbering ch. 2

$j_1(x)$ is the first spherical Bessel function (see equation 2.33).[iii] Thus, in real space, this set of modes exhibits a (decaying) sinusoidal oscillation rather than the sharp real-space top hat used to derive $\delta_{\rm crit}$. The problem is a basic one, ubiquitous in Fourier transforms: we cannot hope to have the simultaneous advantages of real-space top hats (specifically, the simple spherical collapse criterion) and k-space top hats (uncorrelated steps in the random walk). Nevertheless, we may brush this inconsistency aside and assume that the two different filters are "close enough" to be compatible—we use both sharp k-space filtering for the random walks and a sharp top hat to define the critical virialization density. The only justification for such an approach is its eventual success and its simplicity: although one can construct more self-consistent approaches, they ultimately fare no better.

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

$$p_a = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\delta_{\rm crit}(z)}^{\infty} d\delta \, \exp\{-\delta^2/2\sigma^2(M)\}. \tag{3.41}$$

The second class can also easily be computed with a clever "mirror" argument originally due to Chandrasekhar (see Figure 3.11). In fact, these two cases have equal probabilities: any random walk belonging to class *(a)* may be reflected around its first up-crossing of $\delta_{\rm crit}(z)$ to produce a walk of class *(b)*, and vice versa. Hence, the fraction of mass elements included in halos of mass less than M at redshift z is simply

$$f_{\rm coll}(< M) = 1 - 2p_a, \tag{3.42}$$

which may be differentiated to yield the Press-Schechter mass function, equation (3.39). This approach better shows the physical origin of the extra factor of 2 needed to obtain equation (3.38): many of the mass elements may *appear* to be in local underdensities but have actually already been incorporated into larger collapsed halos.

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

[iv] Here we will derive the mass function using a special trick; the general approach to its solution is described later in §9.4.1.

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

The other advantage of the excursion set approach is that it allows us to examine how halos relate to one another other and evolve over time.[10] First, consider how halos at one redshift are related to those at another. Suppose that a halo of mass M_2 exists at redshift z_2. Then, we know that the random function δ_M corresponding to a mass element within the halo first crosses $\delta_{\rm crit}(z_2)$ at k_{c2} corresponding to M_2.

Given this constraint, we may study the distribution of k_c where the function δ_M crosses other thresholds (keep in mind that $\delta_{\rm crit}$ is a function of redshift, so these other thresholds tell us about either the progenitors or "descendants" of the given halo). It is particularly easy to construct the probability distribution for when trajectories first cross some $\delta_{\rm crit}(z_1) > \delta_{\rm crit}(z_2)$ (which implies that $z_1 > z_2$, corresponding to the halo's progenitors); clearly, this must occur at some $k_{c1} > k_{c2}$ or $M_1 < M_2$. Figure 3.11 shows an example of such a pair of objects and thresholds.

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

$$\frac{dP}{dM_1}(M_1, z_1|M_2, z_2) = \sqrt{\frac{2}{\pi}}\left[\frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{\sigma^2(M_1) - \sigma^2(M_2)}\right]\left|\frac{d\sigma(M_1)}{dM_1}\right| \exp\left\{-\frac{[\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)]^2}{2[\sigma^2(M_1) - \sigma^2(M_2)]}\right\}. \quad (3.43)$$

This equation may be rewritten as saying that the quantity

$$\tilde{v} = \frac{\delta_{\rm crit}(z_1) - \delta_{\rm crit}(z_2)}{\sqrt{\sigma^2(M_1) - \sigma^2(M_2)}} \quad (3.44)$$

is distributed as the positive half of a Gaussian with unit variance; equation (3.44) may be inverted to find $M_1(\tilde{v})$.

We can interpret the statistics of these random walks as those of merging and accreting halos. For a single halo, we may imagine that as we go back in time, the object breaks into ever smaller pieces, similar to the branching of a tree. Equation (3.43) provides the distribution of the sizes of these branches at some given earlier time. One can then imagine using this description of the ensemble distribution to generate random realizations of the merger histories of single halos—or "merger trees." One recursively steps back in time, at each step breaking the final object into two pieces and choosing a value from the distribution of equation (3.43) to determine the mass ratio of the two branches.

Unfortunately, this recursion has proven to be difficult in practice. Part of the problem is conceptual: one might want to define "merger rates" by taking the limit of equation (3.43) as $z_2 \to z_1$. However, one immediately finds that the resulting rate is not symmetric in the Press-Schechter theory: the rate

at which objects of mass M merge with objects of mass M' is not equal to the rate at which objects of mass M' merge with objects of mass M! The root of the problem is that, even with the excursion set approach, the Press-Schechter formalism does not divide dark matter particles into discrete objects; rather, it simply computes the statistical properties of the ensemble. Unfortunately, quantities like the merger rate implicitly assume that the objects do sit in discrete objects and ignore smooth accretion of diffuse matter. N-body simulations are the most reliable tool for following the merger statistics.

Nevertheless, we may use the distribution of the ensemble to derive some approximate analytic results that at least provide a helpful guide. A useful example is the distribution of the epoch at which an object that has mass M_2 at redshift z_2 has accumulated half 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:

$$P(z_f > z_1) = \int_{M_2/2}^{M_2} \frac{M_2}{M} \frac{dP}{dM}(M, z_1|M_2, z_2) dM, \tag{3.45}$$

where dP/dM is given in equation (3.43). The factor M_2/M corrects the counting from mass weighting to number weighting; each halo of mass M_2 can have only one progenitor of mass greater than $M_2/2$. Differentiating equation (3.45) with respect to time gives the distribution of formation times. Overall, the excursion set formalism provides a reasonable approximation to more exact numerical simulations of halo assembly and merging histories.

3.4.3 Improvements to the Press-Schechter Formalism

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

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

The second area is a refinement of refine the spherical collapse model itself: as we will see in the next chapter, dark matter structures rarely collapse symmetrically, so it is possible to improve the threshold density $\delta_{\rm crit}(z)$ by including a more accurate physical description. The best motivated such approach is to allow ellipsoidal collapse, in which the three axes of the object collapse at different times. The torques driving this collapse are set by the halo's environment, which depends on the halo mass itself (as will be shown in §3.5). This means that the collapse threshold $\delta_{\rm crit}^{\rm ST}$ is a function of not only redshift but also halo mass, so the absorbing barrier in the diffusion problem is no longer a constant. In particular, the threshold increases as halos get smaller: this increases the

abundance of massive halos and decreases the abundance of small halos, just as needed.

However, the match to numerical simulations is still not perfect, so it is now most common to use a fit to these results; fortunately, detailed simulations show that the resulting function can still *nearly* be phrased as a function of ν_c. Fits of the form[11]

$$n_{\rm ST}(M) = A'\sqrt{\frac{2a'}{\pi}}\frac{\rho_m}{M}\frac{-d(\ln\sigma)}{dM}\,\nu_c\left[1+\frac{1}{(a'\nu_c^2)^{q'}}\right]e^{-a'\nu_c^2/2}\,, \tag{3.46}$$

which is closely motivated by ellipsoidal collapse, perform reasonably well; this is known as the *Sheth-Tormen model*. A reasonably good fit to simulations can be obtained by setting $a' = 0.707$ and $q' = 0.3$ and ensuring the proper normalization by adopting $A' = 0.322$.[12] The parameters here are motivated by ellipsoidal collapse; however, the fit to numerical simulations can be improved by varying them slightly or introducing refined functional forms. For example, at very high redshifts, the form of equation (3.10) can overestimate the abundance of very massive, rare halos by $< 50\%$.[13]

However, recent refined numerical simulations show that even more complex fits, in which the universal dependence on ν_c breaks slightly, are necessary for high precision work, with the fitting parameters depending on redshift.[14] The functions that best fit numerical simulations continue to evolve somewhat as those simulations improve—especially because the high redshift case is relatively unexplored thus far—so we will use the simple form of equation (3.46) for the calculations in this book. Results for the associated comoving density of halos of different masses at different redshifts are shown in Figure 3.10.

Figure 3.12 also shows $f_{\rm coll}$, or the fraction of mass above a given threshold (here shown as a function of $T_{\rm vir}$). The solid curves take the improved mass function in equation (3.46), while the dotted curves take the simpler (but less accurate) form of equation (3.38). The three sets of curves show the fraction in halos with $T_{\rm vir} > 10^3$, 10^4, and 10^5 K. We will see later that the middle value here corresponds to the threshold for efficient star formation before reionization, while the last is approximately the threshold for star formation after reionization. Note that in all cases, $f_{\rm coll}$ increases extremely rapidly at high redshifts, since (at least at $z > 10$) all these halos are far out on the exponential tail of the mass function. Also note that the simple Press-Schechter prescription tends to underestimate the abundance of high-mass halos—and thus drastically underestimates $f_{\rm coll}$ when halos are rare—but slightly overestimates the abundance of low-mass halos (visible in the $T_{\rm vir} > 10^3$ K curves near $z \sim 5$).

3.5 Halo Clustering in Linear Theory

To this point, we have computed the *average* abundance of halos throughout the universe. But, of course, the universe is not perfectly smooth on larger scales, and we naturally expect large-scale overdensities to have an overabundance of halos relative to the average, and large-scale underdensities (or voids) to have

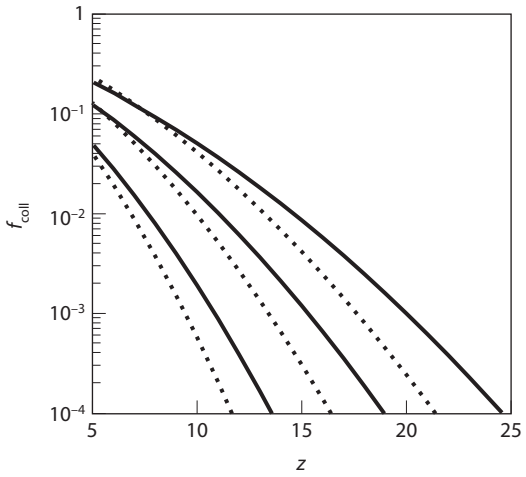


Figure 3.12 The collapse fraction of dark matter halos at high redshifts. The solid curves show $f_{\rm coll}$ computed from the mass function of equation (3.46), which is motivated by ellipsoidal collapse with parameters determined by a fit to numerical simulations. From top to bottom the three curves show the fraction of mass in halos with $T_{\rm vir} > 10^3$, 10^4, and 10^5 K. The dotted curves use the original Press-Schechter form in equation (3.38).

a deficit. This clustering is an extremely important aspect of halos in the real universe, especially at high redshift.

The excursion set formalism allows us to describe this clustering in detail, at least to linear order,[15] and behaves (similarly to a region carved out from a universe with a higher value of Ω_m). A large-scale overdensity corresponds to $\delta_b > 0$ across a large but finite mass M_b. We then imagine our halos (with mass $M < M_b$) forming out of this material. We can solve the related diffusion problem just as for the conditional mass function simply by changing the origin of our random walks from $\delta = 0$ and $M \to \infty$ (or $\sigma^2 \to 0$) to δ_b at M_b (the boundaries of our region). The small head start these modified initial conditions provide halos in overdense regions can be extremely important: recall that the density distribution itself is Gaussian, so the abundance of rare halos is exponentially sensitive to the underlying density (see Figure 3.13). We now wish to describe this dependence, often called the "peak-background split," quantitatively.

transpose; delete parentheses and extra period / PE

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

$$\left(\frac{\delta n}{n}\right)_{\rm halo} = \frac{V}{V(1-\delta_b)} - 1 = \delta_b. \qquad (3.47)$$

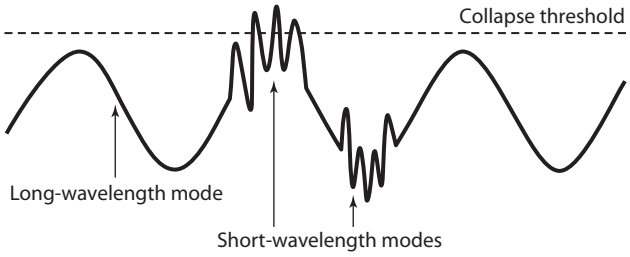


Figure 3.13 Modulation of dark matter halos by the underlying density field. Here, a single long-wavelength mode is shown for simplicity. The dashed horizontal line represents the density threshold for collapse (as in the Press-Schechter model); the large-scale mode does not reach the threshold anywhere. However (to linear order), the superposed small-scale modes evolve independently of the amplitude of this large-scale field. Where the large-scale field has a high overdensity, it is much easier for the small-scale modes to pass the collapse threshold. We therefore expect dark matter halos to be highly clustered inside large-scale overdense regions.

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

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

$$\delta_{\rm crit} \to \delta_{\rm crit} - \delta_b \qquad \text{or} \qquad \nu_c \to (\delta_{\rm crit} - \delta_b)/\sigma(M). \tag{3.48}$$

Note here that, as is conventional in the excursion set approach, we will extrapolate $\delta_{\rm crit}$ at the redshift of interest to the present day by dividing by the growth factor (recall that this makes σ^2 redshift independent). We must therefore also (linearly) extrapolate the background density to the present; this is $\delta_b^0 = \delta_b(z)/D(z)$.

We can therefore immediately write the halo abundance in the region. However, it is most useful to consider a small overdensity $\delta_b^0 \ll \delta_{\rm crit}$ and Taylor expand about the average result. Expanding in a Taylor series,

$$\frac{dn}{dM}(\delta_b^{\rm ex}) \approx \frac{dn}{dM} + \frac{dn}{dM\,d\nu_c}\frac{d\nu_c}{d\delta_b^0}\delta_b^0 + \ldots, \tag{3.49}$$

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

$$\left.\frac{\delta n}{n}\right|_{\rm PS} = \frac{\nu_c^2 - 1}{\sigma\,\nu_c}\delta_b^0. \tag{3.50}$$

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

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

superscript should be 0 / AA (zero)

greek delta subscript roman crit / AA

Combining this effect with the change in volume (equation 3.47), we find

$$\frac{dn}{dM}(\delta_b) = \frac{dn}{dM}[1 + b_{\rm PS}(m)\delta_b], \tag{3.52}$$

where

$$b_{\rm PS}(M) = 1 + \frac{\nu_c^2 - 1}{\delta_c(z=0)}. \tag{3.53}$$

greek delta subscript roman crit / AA

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

We have evaluated the bias for the Press-Schechter mass function; one can perform a similar exercise with the more accurate mass functions described in §3.4.3. For example, the mass function of equation (3.46) yields

greek delta subscript roman crit x2 / AA

$$b_{\rm ST}(M) = 1 + \frac{a'\nu_c^2 - 1}{\delta_c(z=0)} + \frac{2q'/\delta_c(z=0)}{1 + (a'\nu_c^2)^{q'}}. \tag{3.54}$$

This result has the same qualitative trends as the earlier expression, although massive halos tend to be somewhat less clustered and small halos somewhat more. Figure 3.14 shows the bias for this model over the same mass and redshift ranges as in Figure 3.10. Note that galaxy-mass halos ($M > 10^8\ M_\odot$) can be quite highly biased during the era of the first galaxies, while very small halos are "antibiased" ($b_{\rm ST} < 1$) at late times. These halos tend to form in underdense regions, because those in overdense regions have already been swallowed by larger halos. In other words, we see that massive, rare halos tend to live in overdense regions, with many neighbors, while halos below a characteristic mass scale tend to form in isolation. This in turns means that, on average, massive halos grow rapidly by merging with their many neighbors, while low-mass halos will grow more sedately.

3.6 The Nonlinear Power Spectra of Dark Matter and Galaxies

DES: 2 lines under stacked heads ok?

3.6.1 The Halo Model

We have now assembled several powerful ingredients to describe the distribution of matter in the Universe: *(i)* the mean abundance of halos as a function of

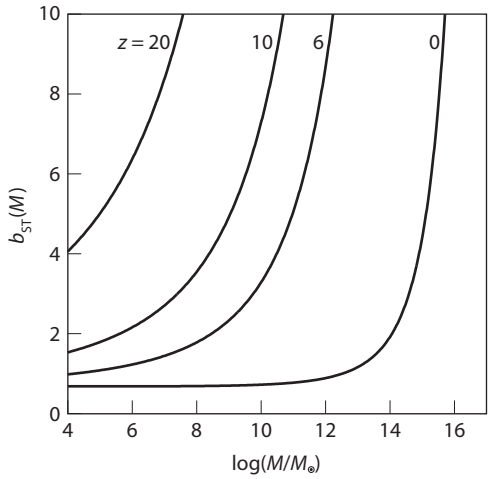


Figure 3.14 The linear bias of halos as a function of M at different redshifts z according to the Sheth-Tormen model (equations 3.46 and 3.54).

mass and redshift; *(ii)* the clustering of these halos; and *(iii)* the radial structure of these halos (the NFW or Einasto profiles). The first two of these ingredients are constructed from linear theory; the third is taken from numerical simulations but is remarkably simple. We can now gather these tools into a first stab at computing the statistical distribution of matter even in the *nonlinear* regime through a powerful approach known as the *halo model*. [16]

The idea is to construct the power spectrum (or, alternatively, the correlation function) of dark matter by conceptually dividing all the matter in the Universe into halos of some—often very small—mass.[v] Because the "universal" halo profile describes the structure of *each* of these halos, and the excursion set formalism describes their abundance and statistical distribution, we can use this picture to compute the correlations between any two dark matter particles (see Figure 3.15).

In the following we use the simpler NFW halo form for concreteness, but the improved Einasto profile is qualitatively similar (the difference is important only on very small scales; see Figure 3.8). We first write the NFW profile for a halo of virial mass M in the simplified form

$$\rho(r|m) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2}, \tag{3.55}$$

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

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

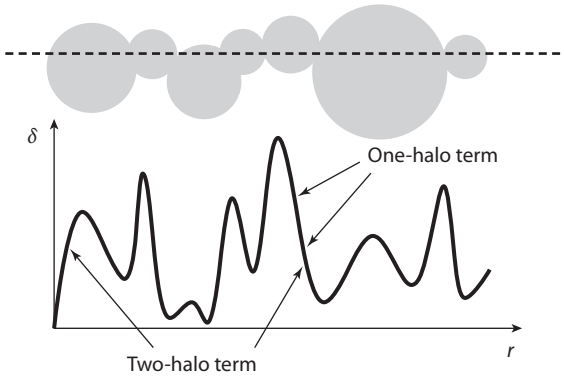


Figure 3.15 Cartoon of the halo model for the cosmic density field. The model splits the dark matter field into gravitationally bound halos (gray circles) with a range of masses. Each of these follows a well-defined density structure, so one can reconstruct the density field purely from the locations of the halos (bottom plot). Correlations among the dark matter particles may then be quantified in two steps: the *one-halo term* describes correlations among particles in a single halo, while the *two-halo term* describes correlations among dark matter particles in separate halos.

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

$$u(k|m) = \frac{4\pi\rho_s r_s^3}{m}\left\{\sin(kr_s)\left[\mathrm{Si}([1+c_\mathrm{N}]kr_s) - \mathrm{Si}(kr_s)\right] - \frac{\sin(c_\mathrm{N}kr_s)}{(1+c_\mathrm{N})kr_s} + \cos(kr_s)[\mathrm{Ci}([1+c_\mathrm{N}]kr_s) - \mathrm{Ci}(kr_s)]\right\}, \quad (3.56)$$

where

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}dt, \quad (3.57)$$

$$\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos t}{t}dt. \quad (3.58)$$

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

$$u_\mathrm{g}(k|m) = \exp[-(kr_s)^2/2]. \quad (3.59)$$

This value is near unity for $k \ll 1/r_s$ before falling off at larger wavenumbers; the shape of any realistic density profile is qualitatively similar. Figure 3.16 shows these halo profiles (both in the NFW and Gaussian forms) for a range of halo masses at $z = 6$. Note that u is flat for $k < r_\mathrm{vir}$ in both cases and then falls off steeply. The rate of decline depends on the halo structure and is somewhat gentler with the power-law density profile of the NFW model (the oscillations are due to the finite size of the halo). Note that the concentration of the larger halos is near unity, so $r_s \approx r_\mathrm{vir}$; it is ∼7 for the smallest halo.

insert superscript -1 after r / AA

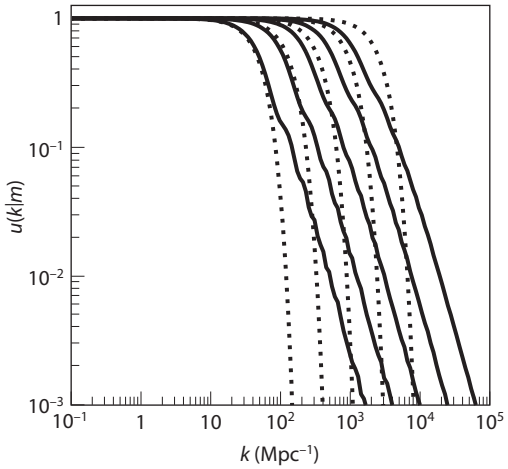


Figure 3.16 Fourier transforms of the normalized halo density profiles for a series of halos at $z = 6$. The solid lines show the profiles for halos following the NFW form and $m = 10^6$, 10^7 10^8, 10^9, and $10^{10}\,M_\odot$, from right to left (equation 3.56). The dotted lines show the Gaussian profiles of equation (3.59), with the same value of r_s.

the / PE

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

$$\rho(\mathbf{x}) = \sum_i \rho(\mathbf{x} - \mathbf{x}_i|m_i) \tag{3.60}$$

$$= \sum_i m_i u(\mathbf{x} - \mathbf{x}_i|m_i) \tag{3.61}$$

$$= \sum_i \int dm \int d^3x' \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) m u(\mathbf{x} - \mathbf{x}'|m), \tag{3.62}$$

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

Now, note the useful identity

$$\left\langle \sum_i \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) \right\rangle = n(m). \tag{3.63}$$

This identity arises because the Dirac delta functions in each volume appear a number of times equal to the number of halos (at each mass) per unit volume.

Thus, the mean density is

$$\bar{\rho} = \int dm \int d^3x' \left\langle \sum_i \delta(m - m_i)\delta(\mathbf{x}' - \mathbf{x}_i) \right\rangle mu(\mathbf{x} - \mathbf{x}'|m) \tag{3.64}$$

$$= \int dm\, mn(m). \tag{3.65}$$

This result is then just the number density of halos multiplied by their mass: recall that we assume that *all* dark matter particles are bound into halos of one mass or another.

3.6.2 The Correlation Function

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

$$\left\langle \sum_{i,j} \delta(m_1 - m_i)\delta(\mathbf{x}_1 - \mathbf{x}_i)\delta(m_2 - m_j)\delta(\mathbf{x}_2 - \mathbf{x}_j) \right\rangle, \tag{3.66}$$

where i and j label the halo sums at the two spatial points. Although analogous to the average in $\bar{\rho}$, this expression is much more complicated, and to evaluate it we must split it into two components, motivated by our picture showing all particles sitting inside halos (see Figure 3.15). In that model, the two particles whose correlation we seek can have two qualitatively different configurations. First, they can sit inside the same halo (so that the indices are the same, $i = j$), in which case the halo density profile uniquely fixes their correlation strength. Here we are summing twice over the same halo, so this part of the average is

$$n(m)\delta(m_1 - m_2)\delta(\mathbf{x}_1 - \mathbf{x}_2) \tag{3.67}$$

In the double integral, we therefore have

$$\begin{aligned} &\int dm\, m^2 n(m) \int d^3x_1 \int d^3x_2\, \delta(\mathbf{x}_1 - \mathbf{x}_2) u(\mathbf{x} - \mathbf{x}_1|m) u(\mathbf{x}' - \mathbf{x}_2|m) \\ &= \int dm\, m^2 n(m) \int d^3x_1 u(\mathbf{x} - \mathbf{x}_1|m) u(\mathbf{x}' - \mathbf{x}_1|m) \\ &= \int dm\, m^2 n(m) \int d^3y\, u(\mathbf{y}|m) u(\mathbf{y} + \mathbf{x}' - \mathbf{x}|m) \equiv \bar{\rho}^2 \xi_{1h}(\mathbf{x} - \mathbf{x}') \end{aligned} \tag{3.68}$$

insert comma / AA

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

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

$$n(m_1)n(m_2)[1 + \xi_{hh}(\mathbf{x}_1 - \mathbf{x}_2|m_1, m_2)] \tag{3.69}$$

where ξ_{hh} measures the correlations between the halos themselves. Fortunately, we can easily compute this value, at least when linear theory applies: we know the linear dark matter power spectrum and hence correlation function $\xi(r)$, and we know from §3.5 how the halo number densities reflect the underlying dark matter density. Therefore,

$$\xi_{hh}(\mathbf{x}-\mathbf{x}'|m_1,m_2) = b(m_1)b(m_2)\xi_{\rm lin}(\mathbf{x}-\mathbf{x}'). \tag{3.70}$$

Note, however, that although equation (3.69) is general, equation (3.70) assumes that fluctuations in the halo distribution remain linear. This is not necessarily the case at high redshifts: even though the dark matter density fluctuations are very small, the halos can be so biased that the *halo* fluctuations are nonlinear (see Figure 3.14). One must be cautious with using linear theory in this regime.

In any case, these off-diagonal terms become

$$\begin{aligned}&\int dm_1\, m_1 n(m_1) \int dm_2\, m_2 n(m_2) \int d^3x_1 \int d^3x_2\, u(\mathbf{x}-\mathbf{x}_1|m_1)u(\mathbf{x}'-\mathbf{x}_2|m_2)\\ &\quad\times[1+\xi_{hh}(\mathbf{x}_1-\mathbf{x}_2|m_1,m_2)]\\ &\quad\equiv \bar\rho^2[1+\xi_{2h}(\mathbf{x}-\mathbf{x}')],\end{aligned} \tag{3.71}$$

where the *two-halo correlation function* ξ_{2h} describes correlations between particles in different halos. For some physical insight, suppose that halos are sharply peaked compared with the separation of interest, or $|\mathbf{x}-\mathbf{x}'| \gg r_{\rm vir}$. Then, we can approximate the profiles as delta functions, and the integrals over x are easy. We therefore obtain (in the linear regime of equation 3.70)

$$\xi_{2h}(\mathbf{x}-\mathbf{x}') \approx \xi(\mathbf{x}-\mathbf{x}') \int dm_1\, \frac{m_1}{\bar\rho} b(m_1)n(m_1) \int dm_2\, \frac{m_2}{\bar\rho} b(m_2)n(m_2). \tag{3.72}$$

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

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

$$\int dm\, \frac{m}{\bar\rho}\left[1+\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] n(m) = 1 + \int dm\, \frac{m}{\bar\rho}\left[\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] n(m) \tag{3.73}$$

$$= 1 + \sqrt{\frac{2}{\pi}} \int d\nu_c \left[\frac{\nu_c^2-1}{\delta_{\rm crit}(z=0)}\right] e^{-\nu^2/2} \tag{3.74}$$

$$= 1, \tag{3.75}$$

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

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

$$\xi(\mathbf{x}-\mathbf{x}') = \xi_{1h}(\mathbf{x}-\mathbf{x}') + \xi_{2h}(\mathbf{x}-\mathbf{x}'). \tag{3.76}$$

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

3.6.3 The Power Spectrum

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

$$P(k) = P_{1h}(k) + P_{2h}(k). \tag{3.77}$$

The one-halo term is straightforward, since it is a simple convolution:

$$P_{1h}(k) = \int dm\, \frac{m^2 n(m)}{\bar{\rho}^2} |u(k|m)|^2. \tag{3.78}$$

The two-halo term is not so trivial. For simplicity, we focus on the linear case in which equation (3.70) applies. First, note that it is a function only of the separation, so we let $\mathbf{x} = 0$ and write (suppressing the integrals over mass and bias for now)

$$\xi_{2h}(\mathbf{r}) \propto \int d^3x_1 \int d^3x_2 u(-\mathbf{x}_1|m_1) u(\mathbf{r}-\mathbf{x}_2|m_2) \xi_{\rm lin}(\mathbf{r}) \tag{3.79}$$

$$= \int \frac{d^3k_1}{(2\pi)^3} u(\mathbf{k}_1|m_1) \int \frac{d^3k_2}{(2\pi)^3} u(\mathbf{k}_2|m_2) \int \frac{d^3k_3}{(2\pi)^3} P_{\rm lin}(\mathbf{k}_3) e^{i\mathbf{k}_2\cdot\mathbf{r}} \times \int d^3x_1 e^{i\mathbf{x}_1\cdot(\mathbf{k}_1+\mathbf{k}_3)} \int d^3x_2 e^{-i\mathbf{x}_2\cdot(\mathbf{k}_2+\mathbf{k}_3)} \tag{3.80}$$

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

In the second line, we took the Fourier transform of each piece and collected exponentials, and in the last line we note that the integrals over $e^{i\mathbf{x}\cdot\mathbf{k}}$ are Dirac delta functions (multiplied by $[2\pi]^3$). Inserting the mass integrals again, we have

$$P_{2h}(k) = P_{\rm lin}(k) \left[\int dm\, \frac{m}{\bar{\rho}} b(m) n(m) u(k|m) \right]^2, \tag{3.82}$$

where $P_{\rm lin}(k)$ is the linear theory power spectrum. Of course, when ξ_{hh} cannot be written simply according to equation (3.70), the expression for P_{2h} is more complex, although the general form of equation (3.77) still applies.

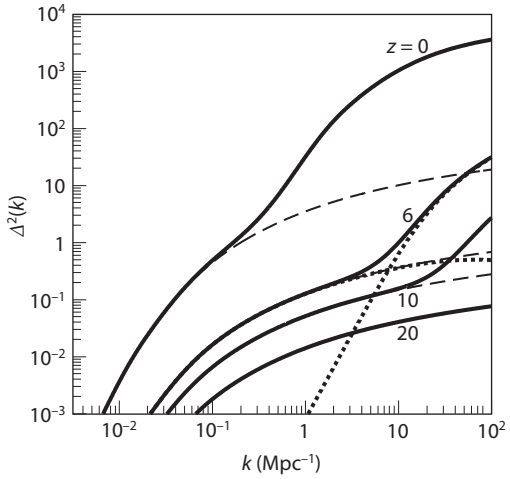


Figure 3.17 Dark matter density power spectrum predictions over a range of redshifts in the context of the halo model. The solid curves show the halo model prediction including only linear bias, while the dashed curves show the corresponding linear theory predictions. For the $z = 6$ curves, we also show separately the one-halo and two-halo terms with the dotted curves; these dominate at large and at small k, respectively.

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

Figure 3.17 shows the resulting power spectra predictions at a range of redshifts as well as a comparison with the underlying $P_{\rm lin}(k)$, including the linear bias approximation. Not surprisingly, on sufficiently large scales the halo model prediction matches $P_{\rm lin}(k)$ precisely; on scales much larger than the halo size, $u(k|m) \to 1$. The factor in brackets in equation (3.82) then becomes the mass-averaged bias, which is just unity, and so $P_{2h} \approx P_{\rm lin}$. Meanwhile, on these scales $P_{1h} \approx$ constant is small.[vi] At large k, the one-halo term—which describes the structures within each halo—dominates; it becomes more and more important as halos grow over time.

Unfortunately, in comparison with detailed numerical simulations at $z > 6$, the halo model prediction is not particularly accurate on the "crossover" scales between the one-halo and two-halo terms at high redshifts; the assumption of linear bias breaks down in this regime. Although the dark matter fluctuations themselves are small at $z > 6$ (see Figure 2.4), massive halos are very highly biased (Figure 3.14). For example, at $z = 10$ the typical (linear theory) density

[vi] Note that this is not universally true: in some applications of the galaxy power spectrum (see §3.64), the constant value of P_{1h} may not be small when the objects of interest are very rare.

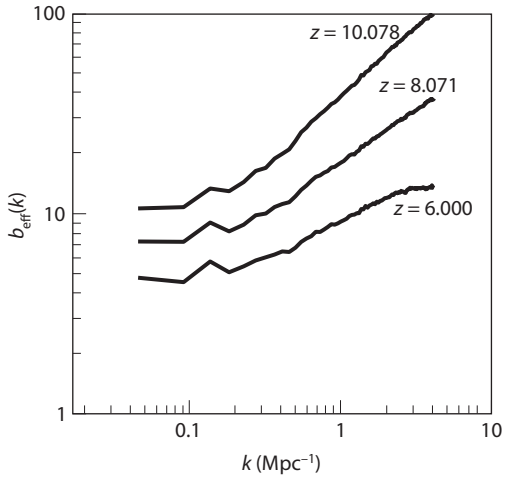


Figure 3.18 Average effective bias for halos with $M > 2 \times 10^9\ M_\odot$ as a function of comoving spatial scale, estimated by comparing the nonlinear and linear power spectra via $b^2_{\rm eff}(k) = P(k)/P_{\rm lin}(k)$, where the "shot noise" term has already been subtracted to obtain the clustering-driven component of the power spectrum $P(k)$. The three different curves vary the redshift of the bias measurement. Fernandez, E. et al., *Astrophys. J.* **710**, 1089 (2010). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

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

One way to account for these nonlinear effects is to continue the expansion of equation (3.49) to higher order in δ_b. However, that result is not easily incorporated into the halo model and the expansion becomes unwieldy once $b\sigma \sim 1$.

A more empirical approach is to allow for an "effective" scale-dependent bias $b_{\rm eff}(k|m)$, which can be measured by comparison with numerical simulations. This function approaches the standard linear bias at small k but increases monotonically toward large k to reflect the nonlinear clustering of nearby halos. Figure 3.18 illustrates this approach from one numerical simulation of high-redshift structure formation. The curves show the average effective bias for halos with $M > 2 \times 10^9\ M_\odot$ as a function of spatial scale, estimated by comparing the nonlinear and linear power spectra via $b^2_{\rm eff}(k) = P(k)/P_{\rm lin}(k)$. Note how they approach constant values (corresponding to the linear theory estimate) at $k < 0.1\ {\rm Mpc}^{-1}$ but then increase rapidly toward smaller scales. Comparison with Figure 3.16 shows that these scales are still much larger than the halos themselves: the culprit is the nonlinear clustering of these very massive halos. In practice, the effect on the total matter power spectrum is smaller, because these halos are so unusual.

3.6.4 The Galaxy Power Spectrum

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

This condition necessitates just a few simple modifications to the preceding formalism. For example, the two-halo term (equation 3.82) has a factor of m inside each integral. This factor effectively counts the number of dark matter particles inside each halo (since each has an identical mass). Instead of counting pairs of particles, we care only about pairs of galaxies,

$$P_{2h}^{\rm gal}(k) = P_{\rm lin}(k)\left[\int dm \frac{\langle N|m\rangle}{\bar{n}_{\rm gal}} n(m) b_{\rm eff}(k|m) u_{\rm gal}(k|m)\right]^2, \qquad (3.83)$$

where $\langle N|m\rangle$ is the mean number of galaxies in a dark matter halo of mass m, and we have included the nonlinear bias correction $b_{\rm eff}$. We have also added two other small adjustments: we normalized to the average number density of galaxies in the sample, $\bar{n}_{\rm gal}$, rather than $\bar{\rho}$, and we included the profile of galaxies within the halo, $u_{\rm gal}$, rather than the dark matter density profile u. In the limit in which linear theory is a good approximation, and on large scales $kr_s \ll 1$, this expression is $P_{2h}^{\rm gal} \approx \langle b\rangle_n^2 P_{\rm lin}$, where $\langle b\rangle_n$ is the bias averaged by *number* of galaxies (rather than mass, as in equation 3.82), because we are considering each galaxy as a single marker, regardless of its mass.

Similarly, the one-halo term has a factor of m^2 reflecting the weighting of pairs of particles within that halo. We need only to change that expression to

$$P_{1h}^{\rm gal}(k) = \int dm\, n(m) \frac{\langle N(N-1)|m\rangle}{\bar{n}_{\rm gal}^2} |u_{\rm gal}(k|m)|^2, \qquad (3.84)$$

delete prime / PE

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

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

Satellites constitute the remaining population: they live inside "subhalos" that have not yet merged completely with the primary halo. Numerical simulations at low redshifts show that subhalos typically appear above some other

[vii] This is not a necessary condition of course: one could easily compute clustering statistics weighted by galaxy luminosity, for example. That is done in many applications where the galaxies are not resolved; see §13.2 for an example.

minimum mass $M_{\rm sat} > M_{\rm min}$, and their number N_s increases roughly proportionally to the halo mass.[17]

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

It is somewhat more difficult to compute the one-halo term. At low redshifts, the number of satellites is found to be roughly a Poisson variable (which is reasonable, since merging is a somewhat stochastic process), so that $\langle N_s(N_s-1)\rangle = \langle N_s\rangle^2$. Including the central galaxy as an additional component implies that $\langle N(N-1)\rangle = \langle N\rangle^2 - 1$, which is sub-Poisson at the low-mass end. In the high-redshift limit, where satellites are unimportant, the one-halo term in the galaxy power spectrum disappears, because there is only one galaxy per halo and hence no correlations.

In addition to these terms, which arise because galaxies trace the density field, we must also add in stochastic "shot-noise" fluctuations arising from the discrete nature of galaxies: any such measurement is fundamentally a counting exercise, so we expect Poisson errors in the galaxy number counts to provide an additional source of fluctuations. In a volume V, the variance in the galaxy number counts is therefore $\sim\bar{n}_{\rm gal}V$, so the fractional density fluctuation in a mode with wavenumber k will be $\Delta^2_{\rm shot} \sim 1/\bar{n}_{\rm gal}V \sim k^3/\bar{n}_{\rm gal}$. A more precise derivation (see §10.4.1) shows $P_{\rm shot}(k) = 1/\bar{n}_{\rm gal}$, or

$$\Delta^2_{\rm shot} = \frac{k^3}{2\pi^2\bar{n}_{\rm gal}}. \tag{3.85}$$

insert overbar (over n) / PE

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

3.7 Numerical Simulations of Structure Formation

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

Fortunately, the rapid increase in computing power over the past several decades has enabled numerical calculations to address this challenge. Computers are particularly well suited to this endeavor, because they can easily calculate the simple physical interactions between many particles. Although following the behavior of individual dark matter particles is still not feasible, numerical

simulations can now (as of 2012) follow the dynamics of $\sim 10^{10}$ particles over long periods of cosmic history. The fundamental idea behind cosmological numerical simulations is to discretize the density field $\rho(\mathbf{x})$ into a large number of particles or grid cells and follow their evolution, incorporating as many physical processes (preferably from first principles) as possible. Such simulations allow detailed comparisons of theoretical predictions with observations as well as the study of "emergent phenomena" that depend on the interaction of many physical inputs and so cannot easily be predicted from analytic models. Nevertheless, one must always bear in mind that a numerical simulation is ultimately no better than the physics underlying its component algorithms, and it is crucial to understand those inputs in assessing the the reliability of linking simulation results to observables in the sky.

Close up / PE

Numerical simulations have been instrumental in elucidating large-scale structure, the Lyman-α forest, the formation of the first stars, and a number of other topics that we will discuss. In the remainder of this section, we briefly discuss their most important features and limitations. However, we defer discussion of computational radiative transfer to section 9.6 and focus here on gravitational and gas dynamics.

3.7.1 Gravitational Dynamics: *N*-Body Codes

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

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{u}_i \tag{3.86}$$

$$\frac{d\mathbf{u}_i}{dt} + 2H(t)\mathbf{u}_i = -a^{-2}\nabla\phi, \tag{3.87}$$

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

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

$$\mathbf{x}_i(t_{n+1/2}) = \mathbf{x}_i(t_n) + \mathbf{u}_i(t_n)\Delta t/2, \tag{3.88}$$

$$\mathbf{a}_i(t_{n+1/2}) = \mathbf{a}[\mathbf{x}_i(t_{n+1/2}), t_{n+1/2}], \tag{3.89}$$

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

$$\mathbf{u}_i(t_{n+1}) = \mathbf{u}_i(t_n) + \mathbf{a}_i(t_{n+1/2})\Delta t, \tag{3.90}$$

$$\mathbf{x}_i(t_{n+1}) = \mathbf{x}_i(t_n) + [\mathbf{u}_i(t_n) + \mathbf{u}_i(t_{n+1})]\Delta t/2. \tag{3.91}$$

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

This scheme requires computing

$$-\nabla\phi(\mathbf{x}_i) = -Gm\sum_{j\neq i}\frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|^3}, \tag{3.92}$$

delete period / AA

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

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

There are three effective resolution limits on N-body codes. The first is the particle mass m, which obviously determines the smallest object that can be followed. Typically, $>10^3$ particles are required to estimate the density profile of a virialized halo reliably, and many orders of magnitude more are required to resolve its substructure in detail.

A second limit emerges from the discretization of the density field: the point-mass force calculation in equation (3.92) causes large artificial deflections when particles pass very close to one another. This is unphysical because the material should actually be distributed over larger volumes, which dramatically reduces the peak force. To alleviate this problem, codes introduce a *force-softening*

parameter such that the force scales as $1/(r^2+\epsilon_F^2)$, which limits the maximum resolvable density contrast to $(\bar{\ell}/\epsilon_F)^3$, where $\bar{\ell}$ is the mean particle spacing. Most modern codes take $\bar{\ell}/\epsilon_F \sim 20$–$50$. For PM codes, the force limit is roughly twice the grid spacing, because it depends on the gradient of the potential across that grid.

The final limit comes from the requirement that the numerical integration over time remain stable. Crudely, this requires that the time steps be sufficiently close so that the first-order approximations intrinsic to the integration converge. Equivalently, the series (cf. equations 3.90 and 3.91)

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \mathbf{u}(t_i)\Delta t + \frac{1}{2}\mathbf{a}(t_i + \Delta t/2)\Delta t^2 + \ldots \tag{3.93}$$

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

$$\Delta t = \epsilon_t \frac{\sigma}{|\mathbf{a}|}, \tag{3.94}$$

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

3.7.2 Hydrodynamics: Grid-Based Approaches

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

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{3.95}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla\Phi - \frac{1}{\rho}\nabla p, \tag{3.96}$$

$$\frac{\partial \varepsilon}{\partial t} + \mathbf{v}\cdot\nabla\varepsilon = -\frac{p}{\rho}\nabla\cdot\mathbf{v} + \frac{(\mathcal{H}-\Lambda)}{\rho}. \tag{3.97}$$

These represent the conservation of mass, momentum, and specific energy (per unit mass) ε, respectively. In the last equation, $\mathcal{H}$ and Λ are the radiative heating and cooling rates per unit volume, respectively, and we have ignored any other internal heating mechanisms. Alternatively, the energy equation can be replaced with an equation for the entropy per unit mass s,

$$\rho T\left(\frac{\partial s}{\partial t} + \mathbf{v}\cdot\nabla s\right) = (\mathcal{H}-\Lambda). \tag{3.98}$$

We have written these equations in an *Eulerian* form, in which the spatial coordinate system is fixed. An alternative is a *Lagrangian* approach, in which the

lowercase greek phi / AA

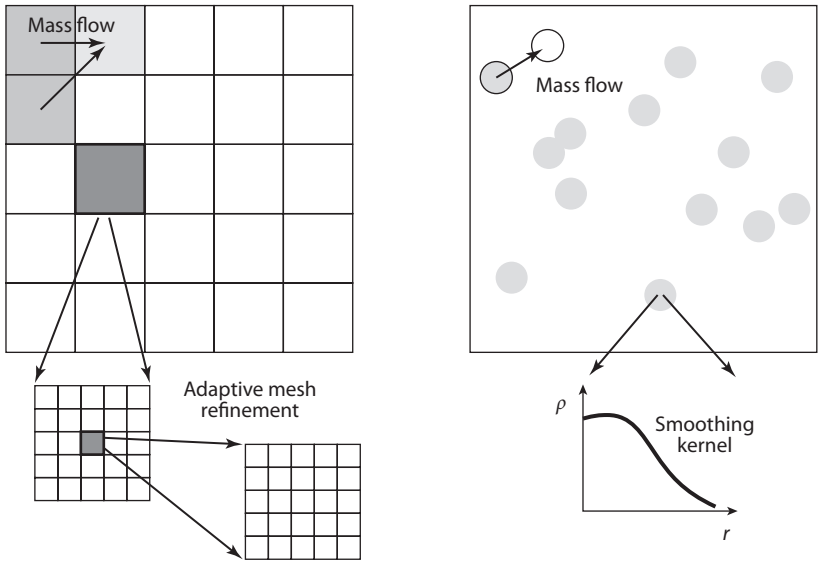


Figure 3.19 Numerical methods for solving hydrodynamics. At left, we show a grid-based algorithm, where the continuous fluid quantities are discretized on a grid, and the equations are solved along the grid faces. Mass flows are described by exchanging material between the cells. Resolution is often increased through *adaptive mesh refinement*, where a new grid is spawned once certain criteria are reached (most commonly, a density or time-step criterion). At right, we illustrate smoothed-particle hydrodynamics, in which the matter field is discretized into a set of particles that can flow freely through the simulation volume. Each particle corresponds to a fixed mass, distributed according to a *smoothing kernel* illustrated at bottom. Mass flows by moving the particles around.

coordinates move with the fluid elements. In this case, the appropriate derivative is the *convective derivative*,

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla, \tag{3.99}$$

so that, for example, equation (3.96) simplifies to

$$\frac{D\mathbf{v}}{Dt} = -\nabla \Phi - \frac{1}{\rho} \nabla p. \tag{3.100}$$

lowercase greek phi / AA

There are two common approaches to solving this system of equations; both are illustrated in Figure 3.19. The first is to divide space into a uniform grid and to solve the hydrodynamic equations for cell-averaged quantities at each grid point. This Eulerian scheme is attractive because the mass, momentum, and energy components of the fluid equations can all be cast as flux conservation

laws

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{F} = 0, \tag{3.101}$$

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

$$q_k(t + \Delta t) = q_k(t) + \frac{\Delta t}{\Delta x} \sum_{j=1}^{3} [F_{j+}(t) - F_{j-}(t)], \tag{3.102}$$

where the indices j label the three axes of the cells, and F_{j+} represents the flux along the j direction between the cell of interest and the next cell in the positive jth direction (and F_{j-} the flux in the negative jth direction). Formulating the fluid equations in this way has one important advantage over the usual differential forms written earlier: when fluids develop sharp discontinuities, like shocks, the latter break down in any scheme with finite resolution. However, the integral forms like equation (3.101), which in the case of shocks are known as the *Rankine-Hugoniot jump conditions*, properly conserve the fluid quantities even if the detailed structures remain unresolved.

The subtlety in grid-based methods lay in ensuring numerical stability for the solutions. For example, the most naive approach to estimating a fluid variable Q at a cell interface is simply to take the midpoint of the cell-averaged quantities in the neighboring cell (e.g., $Q_{k+} = [Q_{k+1} + Q_k]/2$). However, this simple approach is in fact unstable, and more sophisticated algorithms are required. One common strategy is to approximate the calculation as a *Riemann problem* (also known as a shock tube), in which a fluid quantity is constant over two regions with a discontinuity between them. Provided the system obeys conservation laws of the form in equation (3.101), Riemann problems can be solved exactly in terms of *characteristics* that propagate at known speeds in either direction; this exact solution can then be leveraged to calculate the evolution efficiently in more realistic circumstances. (For example, an initially uniform gas with a sharp edge adjacent to vacuum would flow into the vacuum at the sound speed, while a rarefaction wave would travel in the opposite direction through the gas, also at the sound speed.)

One popular technique for leveraging the Riemann problem is known as *Godunov's scheme*. One approximates each cell as a uniform medium at its average value and then solves the Riemann problem at each of its interfaces. The resulting waves can then be propagated into the cell and its new properties calculated at a later time. To avoid collisions and interactions between the waves, the time step must be limited by

$$\Delta t = \epsilon_{\text{grid}} \frac{(\Delta x/2)}{c_s}, \tag{3.103}$$

where c_s is the sound speed, and ϵ_{grid} is a dimensionless constant.

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

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

The fundamental idea of AMR is to demand that the local grid spacing adjust "on the fly" to the physical conditions within the fluid. For example, if a dark matter halo collapses to high density and accretes baryons, the physical resolution must increase to follow the flow. Meanwhile, the time step required with a smaller grid will shrink dramatically according to equation (3.103). AMR codes spawn smaller meshes that are stepped at higher rates, while the background grid continues its slow evolution in low-density regions. While AMR does allow a dramatic increase in the dynamic range of grid-based calculations, the spawning of grids is an imperfect process that leads to some subtle numerical problems when the resolution increases discontinuously, for example in populating the initial conditions of small-scale modes originally absent from the parent grid.

Although AMR solves the most glaring problem with grid-based approaches in astrophysics, these codes suffer from some other important shortcomings. Foremost among them is the violation of Galilean invariance inherent in such methods (i.e., the results of the calculation depend on the reference frame): because the advection terms in equations (3.95)–(3.97) are modeled explicitly, they inevitably contain numerical errors that depend on the magnitude of the bulk velocity relative to the velocity dispersion (which can be large—for example, galaxies merge at velocities comparable to or greater than their own velocity dispersions). This creates numerical viscosity and diffusion that violate Galilean invariance. Without large physical transport coefficients, these numerical artifacts are in fact the leading order terms, so even the qualitative solutions may be questionable under some circumstances. In general, some amount of dissipation is helpful, but limited resolution or high-bulk velocities will cause over mixing. Similar artifacts also appear whenever the bulk velocity is much larger than the thermal velocities; these can be remedied (but not entirely removed) with a careful choice of the reference frame.

overmixing should be one word / PE

3.7.3 Hydrodynamics: Particle-Based Methods

The alternative to grid-based approaches, *smoothed-particle hydrodynamics (SPH)*, discretizes the fluid field and implicitly adapts the resolution to the local fluid properties (see Figure 3.19). SPH is more naturally suited to problems with a high dynamic range of density, but it also faces its own set of challenges.

SPH methods formally aim to recover a smoothed version Q_s of a fluid field Q (such as density or temperature),

$$Q_s(\mathbf{r}) \equiv \int d^3\mathbf{r}'\, Q(\mathbf{r}) W(\mathbf{r}-\mathbf{r}', h), \tag{3.104}$$

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

$$w(x) = \frac{8}{\pi}\begin{cases} 1-6x^2+6x^3, & 0 \le x \le 1/2, \\ 2(1-x)^3, & 1/2 \le x \le 1, \\ 0, & 1 < x. \end{cases} \tag{3.105}$$

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

Now, suppose that we know the fluid properties at a set of points $\mathbf{r}_i$. We associate particles with each of these points and assign mass m_i so as to conserve the total mass in the field and densities ρ_i such that the volume between the particles is $\sim m_i/\rho_i$. We can then estimate the smoothed field Q_s by summing over these particles, so that equation (3.104) becomes

$$Q_s(\mathbf{r}) \approx \sum_k \frac{m_k}{\rho_k} Q(\mathbf{r}_k) W(\mathbf{r}-\mathbf{r}_k, h). \tag{3.106}$$

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

$$\rho_i = \sum_{k=1}^{N_{\rm ngb}} m_k W(\mathbf{r}_i - \mathbf{r}_k, h_i), \tag{3.107}$$

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

Equation (3.106) is generally taken as the SPH estimate for any fluid field. The derivatives of such a field can then easily be calculated (as they require only the derivatives of the kernel W), and from them one can construct discretized versions of the fluid equations. For example, equation (3.95) becomes

3.96 / AA

$$\frac{D\mathbf{v}_i}{Dt} = -\nabla\phi - \sum_{k=1}^{N_{\rm ngb}} m_k \left(\frac{p_i}{\rho_i^2} + \frac{p_k}{\rho_k^2}\right) \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h), \tag{3.108}$$

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

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

The most popular astrophysical codes therefore take a slightly different approach by noting that the fluid equations (3.95)–(3.97) follow from the Lagrangian

$$\mathcal{L} = \int d^3\mathbf{r}\, \rho \left(\frac{|\mathbf{v}|^2}{2} - \varepsilon \right), \tag{3.109}$$

which itself can easily be discretized,

$$\mathcal{L}_{\rm SPH} = \sum_i \left(\frac{m_i |\mathbf{v}_i|^2}{2} - m_i \varepsilon_i \right), \tag{3.110}$$

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

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

$$\frac{D\mathbf{v}_i}{Dt} = -\nabla\phi - \sum_{k=1}^{N_{\rm ngb}} m_k \left[f_i \frac{p_i}{\rho_i^2} \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h_i) + f_k \frac{p_k}{\rho_k^2} \nabla_i W(\mathbf{r}_i - \mathbf{r}_k, h_k) \right], \tag{3.111}$$

where

$$f_i \equiv \left(1 + \frac{h_i}{3\rho_i} \frac{\partial \rho_i}{\partial h_i} \right)^{-1} \tag{3.112}$$

arises from the constraint. Note the similarity to equation (3.108): this slightly more complicated form implicitly includes the particle accounting without much increased complexity, and the direct derivation from a discretized Lagrangian manifestly conserves linear momentum, angular momentum, and energy.

Although elegant, this approach has one key weakness: namely, by assuming a constant entropy, it does not allow shocks or other forms of dissipation. Equation (3.111) must then be supplemented with an *artificial* viscosity that re introduces these features. Perhaps surprisingly, it is relatively easy to formulate this viscosity in such a way that it generates the proper additional entropy at shocks, so long as the prescribed viscosity conserves momentum and energy. This follows because the shock jump conditions (and hence macroscopic fluid variables) are independent of the transport coefficients such as the viscosity. However, SPH codes cannot resolve the structure of the shock itself unless the

viscosity parameter reflects the microphysics of the gas; typically, shocks are much broader in SPH treatments than in grid-based codes. Another challenge is to ensure that this artificial viscosity does not affect the dynamics in regions outside of shocks.

The SPH approach requires a time integrator; because the fluid has been discretized into particles, the leapfrog methods described in §3.7.1 work equally well here (though note that the irreversibility of most hydrodynamics processes actually means that other methods work as well). The time steps must respect the Courant condition of equation (3.94), but because the hydrodynamics equations also involve spatial derivatives, an additional limit applies as well, with $|v\Delta t|/\Delta r < 1$. This, limit is usually written as

$$\Delta t_i = \epsilon_{\rm SPH}\frac{h_i}{c_{s,i}}, \tag{3.113}$$

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

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

It is also worth noting that although SPH simulations do intrinsically adapt to high-density environments, they cannot "zoom" indefinitely. Once the time step of equation (3.113) becomes is too short—say, in runaway gravitational collapse—it becomes impractical to continue the integration. The problem can be circumvented by creating a *sink particle* that accretes mass (and possibly exerts feedback in some prescribed manner) but whose internal structure is not resolved. This technique is used in simulations of star formation and is an example of *subgrid models* that represent physical processes unresolved by the simulation itself (see §3.7.4 for more discussion of these approaches).

Finally, SPH is ill-suited to problems in which the mixing of different fluids is important (such as diffusion), because the particles are generally not allowed to exchange mass. This drawback has more important ramifications than simply following mass around, however: entropy generation through gas mixing is impossible to follow reliably with standard SPH codes.

Although SPH is by far the most popular particle-based solver, it is not the only approach; the kernel is ultimately used only to partition the fluid field into mass elements, and other schemes to accomplish the same purpose can also be used. For example, one can compute a *Voronoi tesselation* for the volume. This

algorithm assigns a volume to each particle that includes all regions closer to its location than to any other particle, without any overlap between the particles. The same Lagrangian technique described previously works with this modified constraint to write the equations of motion for each particle.

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

3.7.4 The Limits of Numerical Simulations

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

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

e / AA

Foremost among them is the enormous dynamic range required to simulate cosmological volumes from "first principles." Ideally, we would like a simulation that resolves star formation inside dwarf galaxies but also contains a representative volume of the intergalactic structures. We will see in chapter 9 that during cosmological reionization, this requires sampling a volume $> (100\,\mathrm{Mpc})^3$. Meanwhile, star formation occurs down a scale $\sim R_\odot = 2.3 \times 10^{-14}$ Mpc. Covering both at once requires a spatial dynamic range $\sim 10^{16}$ (or 10^{48} in mass!), far beyond the capabilities of even the largest computer clusters today or in the foreseeable future.

Cosmological simulations must therefore inevitably incorporate *subgrid models* to approximate physics unresolved by the simulation. The importance of these prescriptions depends on the dynamic range and goals of the simulation. Most commonly, they parameterize processes inside galaxies, including the following:

- *Star formation:* Cosmological simulations, and even simulations of individual galaxies, are far from being able to resolve star formation—and, as

discussed in later chapters, we are still far from understanding that process even if we could zoom in to very small scales. Simulations must therefore construct a subgrid model for star formation, usually calibrating it to an empirical relation such as the Kennicutt-Schmidt law (see §8.5). The computed star formation rates are therefore no more reliable than the empirical or semianalytic model underlying the simulation.

- *Black hole growth:* An equally difficult problem is the accretion of gas onto black holes, which typically occurs on solar system scales inside the complex environments of galactic nuclei. Without resolving the detailed gas dynamics at the center of galaxies (which can be done in specialized simulations, but not in their cosmological-scale counterparts), it is impossible to determine the accretion rates onto these objects from first principles. It is therefore necessary to impose a subgrid model to track the growth of black holes and quasar activity.
- *Galactic winds and feedback:* We will see in chapter 6 (and §8.7) that feedback is likely ubiquitous in star-forming galaxies and crucial for regulating their star formation rates. The energy and momentum injected into the gas by supernovae and radiation likely prevents much of the gas from cooling into stars and removes material from the galaxy, enriching the intergalactic medium (IGM) with metals. However, these processes are difficult to model even in very high-resolution simulations, and simple prescriptions are usually implemented in cosmological simulations. The free parameters are then calibrated to local observations of feedback on galactic scales.

  Even more difficult to model is feedback from supermassive black holes, which can be very important energetically but has very limited observational constraints. Because the feedback occurs most often at the centers of galaxies, the transport of the energy and momentum through the galaxy is crucial for modeling it effectively. For example, nearby radio galaxies launch powerful jets into the IGM, but it is not clear that these jets couple strongly to their host galaxies. With only a crude physical understanding of these processes, subgrid models that make strong assumptions about the underlying coupling mechanisms (in the form of relativistic and nonrelativistic outflows, radiative heating, radiation pressure, or cosmic rays) are necessary.
- *Clumping:* We will see in chapter 9 that small-scale gas clumping is crucial to understanding reionization, but many cosmological simulations do not resolve the relevant physical scales (especially before reionization, when the Jeans mass is small). Moreover, this small-scale structure will evolve as the IGM temperature and pressure change. Often, a subgrid model is inserted to describe this clumping: it can include the clumping from unresolved filaments and sheets in the cosmic web (see chapter 4) as well as the photoevaporation of collapsed "minihalos" that are unable to form stars because their low virial temperature does not allow the gas to cool further. Some reionization simulations ignore hydrodynamics entirely and impose *all* gas clumping through a simplified prescription.

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

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

Another problem, particularly at high redshifts, is ensuring that the simulation samples a representative volume of the Universe. Typically, this is accomplished by demanding that the largest density modes in the simulation remain well in the linear regime at the time the simulation is ended. For technical reasons (to make a FFT easy, and so that the density field "outside" the box can be represented by the box itself), most cosmological simulations implement *periodic boundary conditions*, in which opposite faces of the box are identified with each other. This forces the mean density of the box to take on the average cosmological value, which at first blush automatically appears to make the box a "representative" volume of the Universe. However, for highly clustered objects (which includes galaxies at very high redshifts), this may be misleading, because even a small density boost in a long-wavelength mode can dramatically affect the halo abundance. For sufficiently rare objects, most such objects may actually lie inside large-scale overdensities; a periodic box at the mean density can therefore not contain a fair sample of these halos. Fortunately, this effect can easily be quantified using the conditional mass function in the excursion set formalism (see §3.4.2).

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

In summary, computer simulations are no more intelligent than their creators, and they rely on the proper input physics to produce reliable answers. Their construction and proper use therefore requires as broad and deep a physical understanding as any other area of theoretical astrophysics. Computers follow the algorithms with which they are programmed, and they are limited

by the approximate subgrid physics that was implemented in them. They are therefore most effective at identifying and understanding so-called emergent phenomena, in which complex systems grow from the interactions of simple systems whose physics can individually be accurately described, or in making high-precision predictions for well-understood phenomena. However, if the input physics is incorrect—if the code uses incorrect initial conditions, or excludes any important physical process—the simulation is no better than an analytic model with similar flaws. A recent example is the recognition that baryonic streaming—nominally a second-order effect and so ignored in cosmological simulations of structure formation— potentially provides a crucial modulation of the collapsed matter field (see §2.2.2 and 3.2.2).

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

Chapter Four

The Intergalactic Medium

4.1 The Cosmic Web

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

One of the great triumphs of modern numerical simulations is their description of the distribution of the intergalactic matter distribution in terms of a *cosmic web* of sheets and filaments separating large voids that are nearly empty of matter (see Figure 4.1). However, the formation of these structures is actually remarkably simple, and it can be understood with a simple extension of linear perturbation theory called the *Zel'dovich approximation*[1] and illustrated in Figure 4.2.

change `and' to `as' /AA

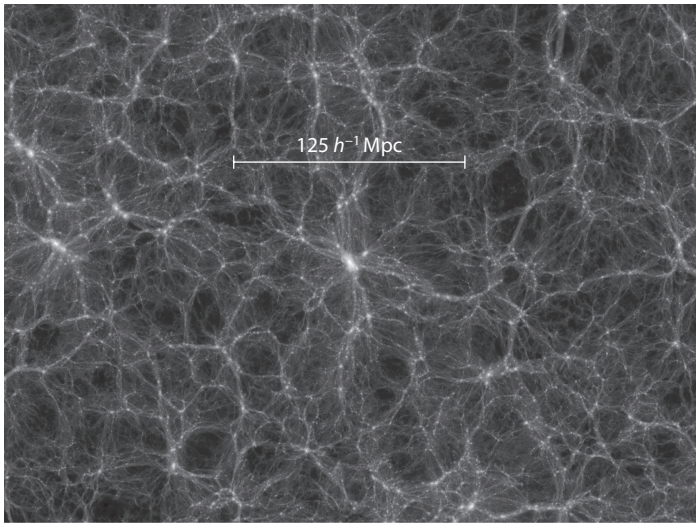
Let us begin by considering the distribution of matter at a very early time t_i. We define $\mathbf{q}$ as the initial comoving position of each particle. If the Universe were homogeneous, we could then write its later position as $\mathbf{r}(t) = a(t)\mathbf{q}$.

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

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

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

The coordinates $\mathbf{q}$ are known as *Lagrangian* coordinates, because they label individual mass parcels; the Lagrangian coordinates of the parcels do not evolve



AU: Add complete reference? This is not from a paper, just supplemental on web.

Courtesy of / AA

Figure 4.1 Slice through the *Millennium Simulation*, a massive computer simulation of cosmological structure formation (see *Color Plate 2* for a color version of this figure). The color scale shows the dark matter density; note how matter is organized into dense filaments (in many cases, these are actually slices through sheets of matter) separating nearly empty voids. Massive galaxies and galaxy clusters form at the intersections of these filaments. V. Springel (2005).

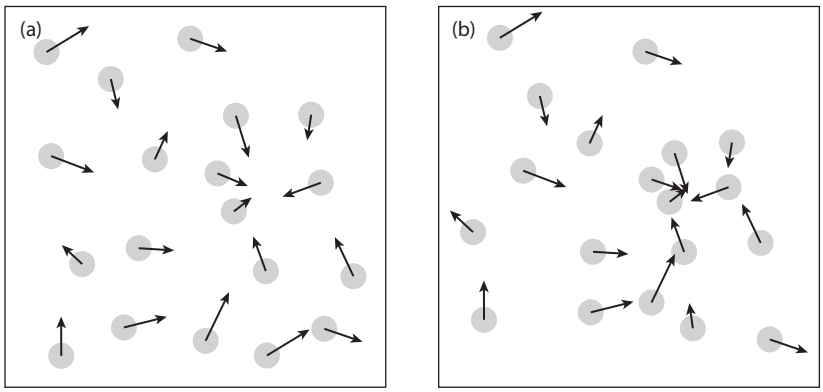


Figure 4.2 Illustration of the Zel'dovich approximation for evolution of the cosmological density field. In panel *(a)*, we show the initial conditions for a calculation (approximating the density field as a set of discrete particles for simplicity). The arrow indicates the direction of the displacement field at the location of each particle. Panel *(b)* shows a later stage in the evolution. In the Zel'dovich approximation, each particle continues to move along the direction of its original displacement (generated from the potential field at the initial time). As a result, matter particles converge into sheets, filaments, and halos.

with time. They are not the same as comoving coordinates $\mathbf{x}$, which are defined by $\mathbf{r}(t) = a(t)\mathbf{x}(t)$. Comoving coordinates are an *Eulerian* system, meaning they refer to a fixed spatial grid rather than labeling particles. The comoving position can evolve as a particle moves, whereas the Lagrangian position $\mathbf{q}$ does not (see also the discussion in §3.7.3).

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

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

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

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

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

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

It is convenient to Fourier transform the density field, as in equation (2.13), except we separate the time dependence of the growing mode:

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

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

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

Not surprisingly, this term has the same form as the peculiar velocity $\mathbf{u}$ in equation (2.15): the displacement field is simply the linear-order peculiar velocity of each particle integrated over time.

By taking the dot product of $\mathbf{k}$ with equation (4.6), we see clearly that $\mathbf{p}(\mathbf{q})$ is the gradient of a function. This implies that the matrix $\partial p_j/\partial q_i$ is a real, symmetric matrix that can be diagonalized to obtain three real eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and their associated principal axes. As a result, the determinant in equation (4.3) may be factored such that

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

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

When $D(t)\lambda_1 = 1$, the collection of particles has collapsed into a sheet perpendicular to the first principal axis. This approximation therefore predicts that two-dimensional "sheets" or "pancakes" are the first nonlinear structures to form. Once collapse occurs along a second axis, a one-dimensional filament forms, and once the third axis collapses, a halo forms.[i] It is this physical picture that motivates the ellipsoidal collapse models used to improve the excursion set approach to halo abundances in §3.4.1.

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

4.2 Lyman-α Absorption in the Intergalactic Medium

Although dark matter dominates the mass budget of the IGM, it is the baryons that most concern us, because they provide the fuel for galaxy formation, interact with the radiation from galaxies, and—most important—provide observables that allow us to trace the structure of the cosmic web. Now we are in a position to study how these baryons are distributed in the Universe.

Hydrogen is the most abundant element in the Universe, making up ≈93% of the atoms in the Universe (the remainder is almost all helium). This predominance of hydrogen is now a well-understood result of the hot Big Bang model, in which nucleosynthesis (completed within the first few minutes after the Big Bang) efficiently combined all the remaining neutrons into helium atoms but then got bottlenecked by the lack of stable isotopes with five or eight nucleons. As a result, all the heavier elements were formed in the interiors of stars within galaxies. We expect (and observations confirm) that the IGM is

[i] Strictly speaking, this complete collapse does not occur in the Zel'dovich approximation, because the particles continue to travel in their original direction of motion. Thus, shortly after collapse to a sheet, the particles cross each other, and the sheet expands again. Obviously, the problem lies in assuming a constant peculiar velocity set by the initial potential; once collapse occurs, the potential has changed significantly. The so-called adhesion model[2] improves the Zel'dovich approximation to account for this effect.

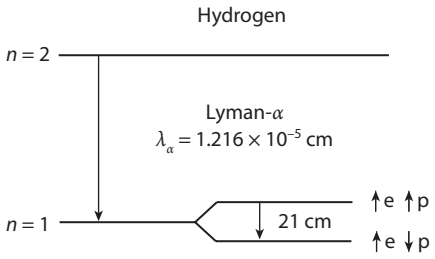


Figure 4.3 Two important transitions of the hydrogen atom. The 21-cm transition of hydrogen occures between the two hyperfine states of the ground energy level (principal quantum number $n = 1$). In the higher energy state, the spin of the electron (e) is aligned with that of the proton (p), and in the lower energy state the two are antialigned. A spin flip of the electron results in the emission of a photon with a wavelength of 21 cm (or a frequency of 1,420 MHz). The second transition is between the $n = 2$ and the $n = 1$ levels, and results 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).

even more dominated by hydrogen and helium than the Milky Way. We therefore focus on these two elements—and especially hydrogen—in our study of that material.

Since the lifetime of energy levels with principal quantum number $n > 1$ is far shorter than the typical time it takes to excite them in the rarefied environments of the Universe, hydrogen is nearly always found to be in its ground state (lowest energy level) with $n = 1$. This implies that we should focus on the transitions that involve the $n = 1$ state. In this book, we describe two such transitions in detail, both depicted in Figure 4.3 (see also chapters 11 and 12).

The most widely discussed transition of hydrogen in cosmology is the Lyman-α spectral line, in which an electron moves between the $n = 1$ and $n = 2$ electronic states and which was discovered experimentally in 1905 by Harvard physicist Theodore Lyman. This line has traditionally been used to probe the ionization state of the IGM in the spectra of quasars, galaxies, and gamma-ray bursts. In 1965, Peter Scheuer[3] and, independently, Jim Gunn and Bruce Peterson[4] realized that the cross section for Lyman-α absorption is so large that the IGM should be opaque to it even if its neutral fraction is as small as $\sim 10^{-5}$.

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

The full cross section of a single atom is

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

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

In practice, the thermal velocities of the IGM atoms have a finite spread. These motions, as well as peculiar velocities and (possibly) turbulence, move the line center around in velocity space (see §11.1.1 for details). However, these velocity shifts are small compared with the cosmological redshift, and we can safely ignore them so long as the photon begins its journey with a wavelength λ much farther from resonance than the line width (see §11.2 for a discussion of the more general case).

We can then approximate the line as narrow,

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

where the prefactor is the integral of equation (4.8) over frequency. Then, if r is the photon's proper distance from the observer, and the neutral hydrogen density is $n_{\rm H\ I}(z) = {\rm x}_{\rm H\ I} n_{\rm H}(z)$, where $x_{\rm H\ I}$ is the neutral fraction, and $n_{\rm H}$ is the number density of hydrogen nuclei,

$$\begin{aligned}
\tau_\alpha &= \int dr\, \sigma_\alpha(r) n_{\rm H\ I}(r) \\
&= \frac{c}{H_0} \int \frac{da}{a} \sigma_\alpha(\nu_{\rm obs}/a) n_{\rm H\ I}(a) [\Omega_m / a^3 + \Omega_\Lambda]^{-1/2} \\
&= \frac{3\Lambda_\alpha \lambda_\alpha^3}{8\pi} \frac{x_{\rm H\ I} n_{\rm H}(z)}{H(z)} \qquad (4.10) \\
&\approx 1.6 \times 10^5 x_{\rm H\ I} (1+\delta) \left(\frac{1+z}{4}\right)^{3/2}, \qquad (4.11)
\end{aligned}$$

where we have used $dr = c\,dt = c\,da/\dot{a} = c\,(da/aH)$ with the Hubble parameter $H = (\dot{a}/a)$ evaluated in the matter-dominated era. We have also let $n_{\rm H}(z) = \bar{n}_{\rm H}(z)(1+\delta)$ in the last line, where $\bar{n}_{\rm H}$ is the mean cosmic density. This *average* IGM optical depth in the Lyman-α transition is referred to as the *Gunn-Peterson optical depth*.

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

In practice, the IGM absorption is observed against a luminous background source (either a bright quasar or bright gamma-ray burst afterglow). The source emits photons over an extended continuum, which allows us to see absorption over a range of wavelengths. If the source resides at a redshift z_s, its Lyman-α transition appears at an observed wavelength $\lambda_\alpha(1+z_s)$. Photons redward of this point begin their journeys at $\lambda > \lambda_\alpha$ and redshift as they travel, so they never enter resonance with the Lyman-α line in the IGM (though they may be absorbed by other species; see §4.6).

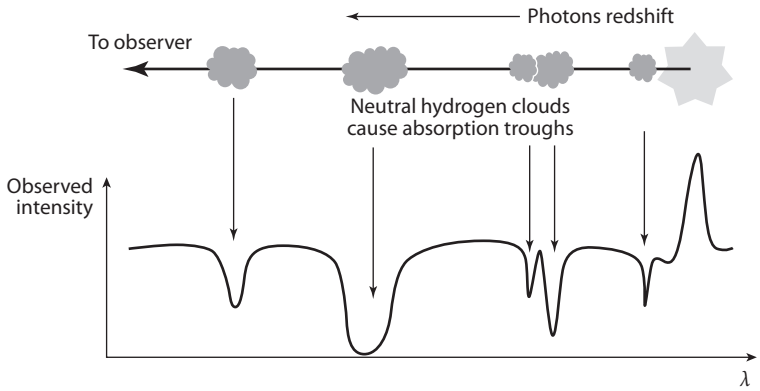


Figure 4.4 The Lyman-α forest. At the top of the panel, light from a distant quasar passes through several clouds of neutral hydrogen along the line of sight. As these photons travel, they redshift. Photons beginning blueward of the Lyman-α transition at the quasar eventually redshift through that resonance. If they do so inside an H I cloud, they may be absorbed by those atoms, creating an absorption feature in the quasar's spectrum. The spectra of distant quasars therefore allow us to map the IGM's density and ionization structure.

Photons blueward of this point eventually redshift into resonance and (if the gas is not too highly ionized) are absorbed. Each such photon redshifts into resonance at a particular distance from the observer (and source) that depends on its initial wavelength: photons emitted far blueward of Lyman-α in the source frame travel a great distance before their wavelength redshifts to 1216 Å, while those emitted just blueward of it reach the resonance near the source. The photon is then absorbed if there is neutral hydrogen at this particular point in the IGM—once it passes that point and redshifts further, it no longer interacts with hydrogen atoms. Thus, each *observed* wavelength samples a different point along the line of sight, and we can map the distribution of H I over a large region along the line of sight to a particular source, as illustrated in Figure 4.4.

The resulting *Lyman-α forest* is so named because of the strong variability of these absorption features. This variability is illustrated in Figure 4.5, which shows three examples of Lyman-α forest spectra at moderately high redshifts ($z \sim 4$). Redward of 1216 Å (in the source frame), the quasar continuum is largely unaffected by the IGM, but blueward of Lyman-α there is highly variable absorption depending on the detailed structure along the line of sight. We now understand this forest of features to originate from the cosmic web: as a line of sight passes through the sheets, filaments, and voids of the cosmic web, the optical depth fluctuates. It is this forest that provides most of our knowledge about the IGM at moderate and low redshifts, and we will next study the physics behind it.

While these features can be identified individually with high-resolution spectra, at lower resolution they blend together as a "trough" of absorption. We

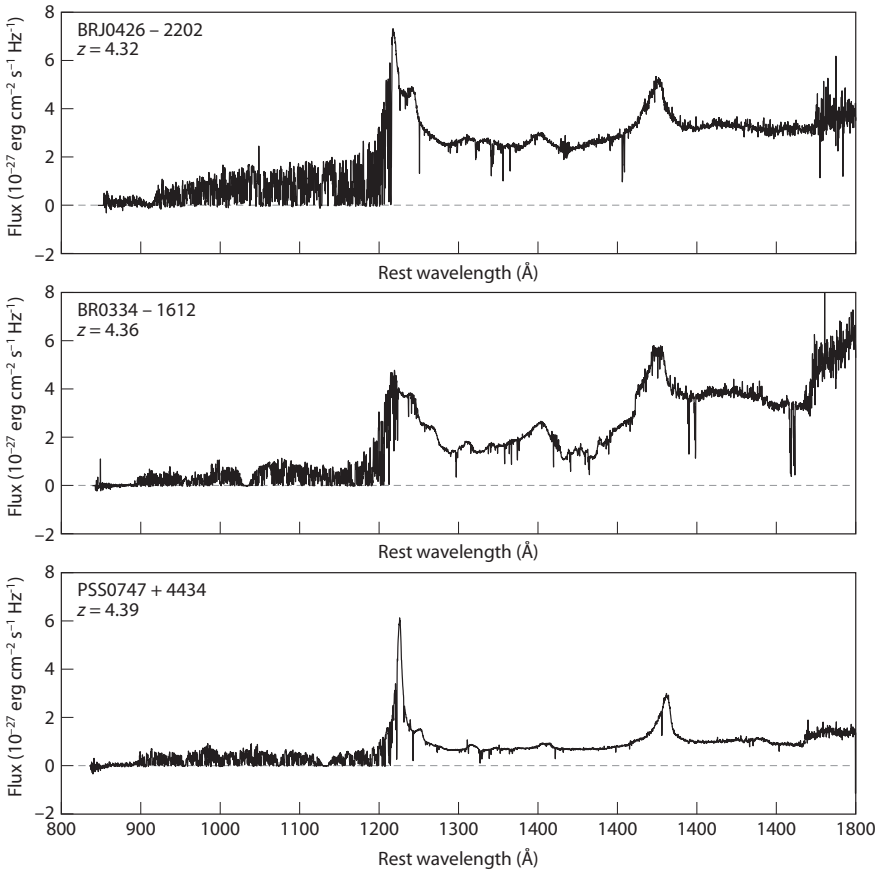


Figure 4.5 Example Lyman-α forest spectra taken with the ESI spectrograph on the Keck telescope. The highly variable absorption blueward of the Lyman-α rest wavelength (at 1216 Å) is the Lyman-α forest, not noise. Redward of this transition, the quasar continuum is visible, with only a few absorption features due to metal line absorbers in the IGM. Songaila, A., & Cowie, L., *Astrophys. J.* **721**, 1448 (2011). Reproduced with permission of the American Astronomical Society.

by / AAS. /

therefore expect a break in the spectrum at $\lambda_\alpha(1+z_s)$, with a depth that depends on the ionized fraction of the IGM and z_s (which affects the proper density of hydrogen). At moderate and high redshifts ($z > 3$), this "Lyman break" is substantial enough to be useful as a redshift estimator. In fact, one of the premier techniques for identifying high-z galaxies is by photometrically searching for extended sources with strong flux redward of the wavelength corresponding to the desired z_s and little or no flux blueward of that wavelength (see §10.2.2).

Naively, how would one expect the optical depth to evolve? The redshift factor $(1+z)^{3/2}$ reflects the evolution of the column density of hydrogen atoms and

implies a slow increase for τ_α as redshift increases. But more important is the factor $x_{\rm H\,I}$, which evolves with both the cosmic density and the ionizing background. As redshift increases, one might naturally expect the number of ionizing sources to decrease, because structure formation is less advanced. In that case we would expect the optical depth to increase even faster than $(1+z)^{3/2}$, and the IGM eventually to become opaque. (In practice, the ionizing background appears to be roughly constant with redshift at $2 < z < 5$, but it must eventually decrease at higher redshifts.)

If a source were to be observed when the atomic fraction of hydrogen was substantial, then *all* photons with wavelengths just short of $1216(1+z_s)$ Å would redshift into resonance, be absorbed by the IGM, and then be reemitted in other directions. Eventually, this process would result in an observed *complete* absorption trough blueward of λ_α in the source spectrum, known as a *Gunn-Peterson trough*.

Figure 4.6 shows spectra of 19 quasars at $z \sim 6$; note how, indeed, the fraction of transmitted flux blueward of the Lyman-α line of each quasar decreases toward the higher redshifts in this range. The spectra of the highest-redshift quasars at $z < 6.4$ show hints of a Gunn-Peterson effect. Unfortunately, this effect is difficult to interpret, because only a very small neutral fraction is required to saturate the Gunn-Peterson trough (see equation 4.11). We cannot yet determine whether the IGM is slightly ionized or nearly neutral at this time; we discuss the Lyman-α line at very high redshifts in §4.7 and again in chapter 11.

Although the Lyman-α transition has so far proved the most useful in understanding the IGM, it is not the only approach. Its key feature is simply that it is a spectral line, so that each observed wavelength corresponds to a different distance from us. Any other spectral line has the same feature and can in principle be used in the same way. Higher Lyman-series transitions are one possibility: they are useful for some applications (see §4.7, for example), but they suffer contamination from the Lyman-α forest and so are more difficult to study. Another possibility is the 21-cm *spin-flip* line, a hyperfine transition of the ground state of H I (see Figure 4.3). The disadvantage of this transition is that it is extremely weak—with an optical depth about seven orders of magnitude smaller than Lyman-α. Thus, it is observable only when the H I density is very large—either because the system itself is very dense or because the IGM is nearly neutral. In either case, the Lyman-α optical depth $\tau_\alpha \gg 1$, so its transmission is extremely small and not detectable for the same regime. The two lines therefore complement each other as probes of the IGM. We will see in chapter 12 that the spin-flip 21-cm transition has great potential for studying the earliest phases of structure formation.

DES: 2 lines under head ok?

4.3 Theoretical Models of the Lyman-α Forest

To compute the optical depth distribution of the IGM it is therefore necessary to know how the neutral fraction $x_{\rm H\,I}$ varies through space. To a very good

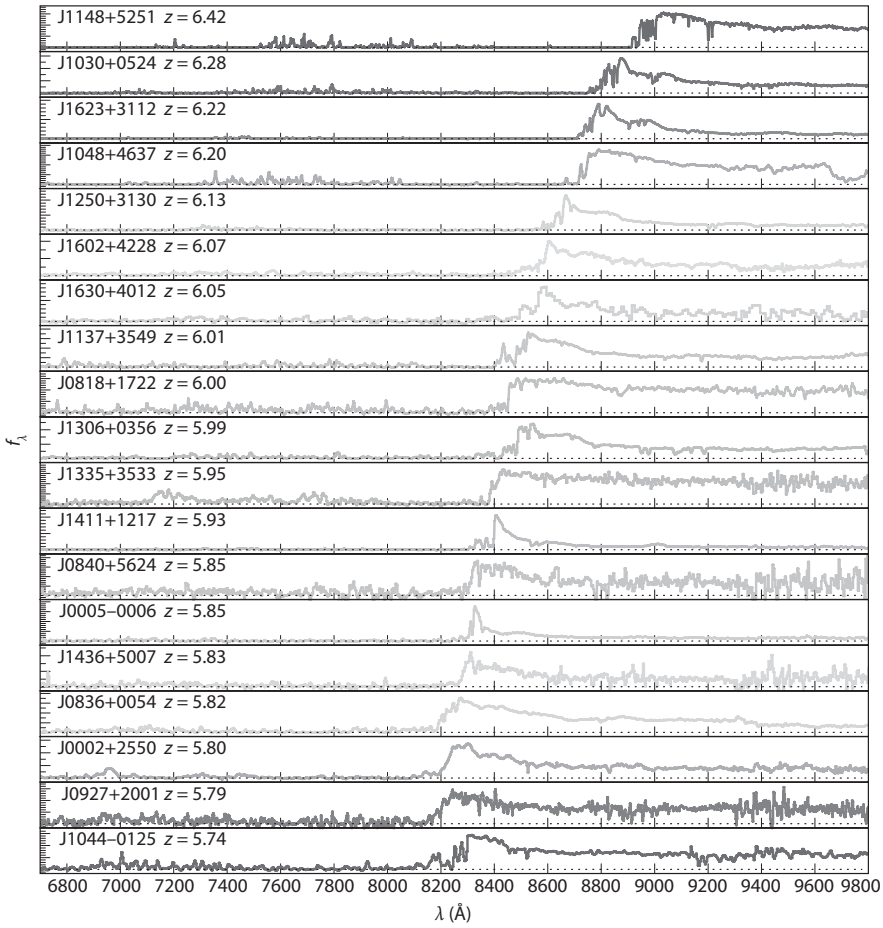


Figure 4.6 Observed spectra of 19 quasars with redshifts $5.74 < z < 6.42$ from the Sloan Digital Sky Survey (see *Color Plate 3* for a color version of this figure). For some of the highest-redshift quasars, the spectra show no transmitted flux shortward of the Lyman-α wavelength at the quasar redshift, 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 broadband photometry is adequate for inferring the redshift of sources during the epoch of reionization. Fan, X., et al., *Astron. J.* 128, 515 (2004). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

approximation, almost all regions are in *ionization equilibrium*, that is, the number of ionizations per second balances the number of recombinations,

$$n_e n_p \alpha(T) = n_{\mathrm{H\,I}} \Gamma \qquad (4.12)$$

132, 117 (2006). (132 is bold, remainder plain) / AA

AU: permission form is for "131, 117 (2006)." Correction needed? ANSWER: It should be 132, 117 (2006).

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

$$\Gamma = \int_{\nu_L}^{\infty} d\nu \, \frac{4\pi J(\nu)\sigma_{\rm H\ I}(\nu)}{h\nu}, \tag{4.13}$$

where $J(\nu)$ is the specific intensity of the background field (in units of erg cm^{-2} s^{-1} Hz^{-1} sr^{-1}), and $\sigma_{\rm H\ I}$ is the cross section for ionization. This integral counts the number of photons per second striking an atom, weighted by the ionization cross section. As we will see, typical values for the ionization rate are $\Gamma \approx 10^{-12}$ s^{-1}, so we normalize $\Gamma = \Gamma_{12} \times 10^{-12}$ s^{-1} for convenience. At this ionization rate, the typical timescale for a neutral atom to be photoionized is $1/\Gamma \sim 10^5$ yr, which is much smaller than $H^{-1}(z)$. Thus, ionization equilibrium is an excellent approximation.

In detail, both n_e and the amplitude of the radiation background depend on the ionization state of helium, which accounts for $\sim 10\%$ of the total electrons in the Universe. We should therefore write two analogs of equation (4.12), for He I and He II, and solve the coupled system of equations given the radiation background. Although this detail is important for precision calculations, it does not qualitatively affect the methods or results, so we usually ignore helium in the following presentation for pedagogical reasons.

The bound-free absorption cross section from the ground state of a hydrogenic ion of species i with nuclear charge Z and an ionization threshold $h\nu_i$ is given by[5]

$$\sigma_{bf}(\nu) = \frac{6.30 \times 10^{-18}}{Z^2} \ {\rm cm}^2 \times \left(\frac{\nu_i}{\nu}\right)^4 \frac{e^{4-(4\tan^{-1}\epsilon_{bf})/\epsilon_{bf}}}{1-e^{-2\pi/\epsilon_{bf}}} \quad \text{for } \nu \geq \nu_i, \tag{4.14}$$

where

$$\epsilon_{bf} \equiv \sqrt{\frac{\nu}{\nu_i} - 1}. \tag{4.15}$$

For neutral hydrogen, $Z = 1$, and $\nu_{\rm H\ I} = (c/\lambda_c) = 3.29 \times 10^{15}$ Hz ($E_{\rm H\ I} = h\nu_{\rm H\ I} = 13.60$ eV); for singly ionized helium, $Z = 2$, and $\nu_{\rm He\ II} = 1.31 \times 10^{16}$ Hz ($E_{\rm He\ II} = h\nu_{\rm He\ II} = 54.42$ eV). Although hardly obvious, equation (4.14) ~~this~~ follows $\sigma_{bf} \propto \nu^{-3}$ near the ionization threshold.

The cross section for neutral helium is more complicated; when averaged over its narrow resonances it can be fitted to an accuracy of a few percent up to $h\nu = 50$ keV by the function[6]

$$\sigma_{bf,{\rm He\ I}}(\nu) = 9.492 \times 10^{-16} \ {\rm cm}^2 \times \left[(x-1)^2 + 4.158\right] \times y^{-1.953}\left(1 + 0.825 y^{1/4}\right)^{-3.188}, \tag{4.16}$$

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

[ii] Here and later we make one subtle simplification by assuming that each photon can ionize only one atom; in reality, the secondary electron liberated during the ionization can then ionize additional atoms. See §9.8.2 for more details.

The radiative recombination coefficient $\alpha(T)$ describes the rate at which electrons and protons recombine (while emitting a photon). Of course, the recombination can occur to any of the hydrogen atom's energy levels; an important special case is recombination to the ground state, which generates a new ionizing photon. Provided that photon is reabsorbed by the gas, such a recombination does not lead to a net increase in the neutral fraction. It is therefore often useful to consider the *case B* recombination coefficient α_B, which excludes recombinations to the ground state. For hydrogen,[7]

$$\alpha_B(T) \approx 2.6 \times 10^{-13} T_4^{-0.76} \text{ cm}^3 \text{ s}^{-1}, \tag{4.17}$$

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

$$\alpha_A(T) \approx 4.2 \times 10^{-13} T_4^{-0.76} \text{ cm}^3 \text{ s}^{-1}. \tag{4.18}$$

Note that both rates are fairly slow in the IGM, except at high redshifts. At the mean density of the IGM, the ratio of the case B recombination time, $t_{\rm rec}^B = 1/n_{\rm H}\alpha$, to the Hubble time t_H is / PE

$$\frac{t_{\rm rec}^B}{t_H} \approx 0.8 \left(\frac{8}{1+z}\right)^{3/2}. \tag{4.19}$$

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

The appropriate coefficient to use depends on the physical situation at hand. If one is concerned with the average absorption in a uniform IGM, case B is clearly the better choice, because photons from recombinations to the ground state are absorbed somewhere else in the IGM. If, however, the IGM is very clumpy so that most of the recombination photons are absorbed inside dense neutral blobs without influencing the low-density IGM about which we principally care, case A is a better choice. Similarly, if one considers ionization equilibrium in a single dense cloud, case A may be more appropriate if the recombinations occur preferentially on the "skin" of the cloud, so that the resulting photons can easily escape to the external medium.

In the highly ionized limit of equation (4.12), we can equate n_p to the total proton density (including those inside hydrogen atoms); in that case, using the case B recombination rate,

$$x_{\rm H\,I} = n_e \alpha_B(T) \Gamma^{-1} \sim 4 \times 10^{-6} (1+\delta) \left(\frac{1+z}{4}\right)^3 T_4^{-0.76} \Gamma_{12}^{-1}, \tag{4.20}$$

where $(1+\delta) = \rho/\bar{\rho}$. Note that for detailed calculations we should include the electrons from ionized helium, but that makes only a minor difference at the level of $\sim 10\%$.

Because observations show that $\Gamma_{12} \sim 1$, we know that the gas is indeed highly ionized, at least at moderate and low redshifts. Conveniently, substituting this value into equation (4.11), the optical depth for gas at the mean density (and $z \sim 3$) is of the of order unity, just in the range in which we can accurately measure the absorption. The Lyman-α forest therefore allows us to map

the cosmic web in exquisite detail, even though the gas itself is at extremely low densities and neutral fractions.

4.3.1 The Temperature–Density Relation

To proceed further and evaluate τ_α as a function of density, we must determine the temperature of the IGM gas. Thermal equilibrium is typically established by three competing effects. The first two are cooling processes: the adiabatic expansion of the Universe and Compton cooling (which is only important at $z > 6$; see §2.2). Other mechanisms—such as line cooling—are much less efficient.

However, photoionization heats the gas. A typical ionizing photon has $h\nu > h\nu_i$, so the free electron is left with some residual kinetic energy. The electron then scatters through the IGM and deposits its energy as heat. The heating rate per *particle* for H I (in K s^{-1}) is[iii]

$$\mathcal{H}_{\rm ph,H\ I} = x_{\rm H\ I} \int_{\nu_{\rm H\ I}}^{\infty} d\nu\, (4\pi J_\nu)\sigma_{\rm H\ I}(\nu) \left(\frac{h\nu - h\nu_{\rm H\ I}}{h\nu}\right). \quad (4.21)$$

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

The temperature of a given parcel of gas therefore evolves following

$$\frac{dT}{dt} = -2HT + \frac{2T}{3}\frac{d\ln(1+\delta)}{dt} - T\frac{d\ln(2-x_{\rm H\ I})}{dt} + \frac{2}{3k_B n_{\rm tot}}(\mathcal{H} - \Lambda), \quad (4.22)$$

where the first two terms account for adiabatic expansion, the third for the change in the total particle density, and the last for radiative heating and cooling: $\mathcal{H} \approx \mathcal{H}_{\rm ph}$ in most cases, and Λ is dominated by Compton cooling (see also §9.8.2). In this expression $n_{\rm tot}$ is the total particle density (including free electrons). We then imagine that the parcel begins as a neutral region and is ionized over some period of time by luminous sources, until it reaches ionization equilibrium with a (slowly evolving) metagalactic background. We would like to understand how its temperature evolves through these two stages.

If we imagine that a gas parcel is initially neutral and then is rapidly exposed to a strong ionizing background, all the gas will quickly be ionized. In this regime, equation (4.22) simplifies substantially, because only the particle number and photoheating terms are large. The final value is then simply the average excess energy per ionization $\langle E_i \rangle$,

$$k_B \Delta T = \frac{2}{3}\frac{n_{\rm H}}{n_{\rm tot}} \langle E_i \rangle . \quad (4.23)$$

This deceptively simple expression actually hides a fair amount of physics in the factor $\langle E_i \rangle$, which depends on how the spectrum of incident radiation interacts with the gas parcel. Two limits are illuminating. First, if the parcel is optically

[iii] Again, here we assume that all the excess energy of the photon goes to heat the gas; in reality, some helps ionize it and some collisionally excites neutral atoms. See §9.8.2.

thin, then the weighting by $\sigma_{\rm H\,I}$ in equation (4.21) reduces the impact of high-energy photons. In this case,

$$\langle E_{i,\rm thin}\rangle \approx \frac{1}{\Gamma}\int_{\nu_i}^{\infty} d\nu\,(4\pi J_\nu)\sigma_{\rm H\,I}(\nu)\left(\frac{h\nu - h\nu_i}{h\nu}\right). \tag{4.24}$$

However, if the element is optically thick up to some maximum frequency (set by the frequency dependence of the cross section, $\sigma_{\rm H\,I} \propto \nu^{-3}$), then *all* photons below this frequency are absorbed, and the weighting by $\sigma_{\rm H\,I}$ disappears.

These two limits can make a significant difference to the total temperature increase: for a specific luminosity $L_\nu \propto \nu^{-\alpha}$, we obtain $\langle E_{i,\rm thin}\rangle = E_{\rm H\,I}/(\alpha+2)$. In the particular case of $L_\nu \propto \nu^{-2}$ appropriate for a low-metallicity galaxy, $\langle E_{i,\rm thin}\rangle / E_{\rm H\,I} \approx 1/4$, if we include all photons between the H I and He II ionization thresholds. In the optically thick limit, we have instead $\langle E_{i,\rm thick}\rangle = E_{\rm H\,I}/(\alpha-1)$. For $L_\nu \propto \nu^{-2}$, this yields $\langle E_{i,\rm thick}\rangle / E_{\rm H\,I} \approx 3/5$ (again including all photons in the range 13.6–54.4 eV). The net temperature change is then $\Delta T \approx 0.5(2/3k_B)\langle E\rangle \sim 30{,}000$ K for the optically thick case, significantly above the value of $\sim 12{,}500$ K for the optically thin case.

Because this energy input is identical for each particle (modulo the optical depth of its environment), the temperature of a parcel should be *independent* of its density immediately after ionization.[iv] However, after this initial phase of ionization, $\mathcal{H}_{\rm ph}$ decreases dramatically, because $x_{\rm H\,I}$ becomes very small.

At this point, the temperature approaches a quasi–steady state, varying only slightly on cosmological scales as the expansion rate changes. The temperature in this state *does* depend on density through the adiabatic cooling rate (the second term on the right-hand side in equation 4.22)—underdense voids can be considered (locally) to have a smaller Ω_m, and so they expand faster. Thus, the low-density regions cool fastest. Numerical calculations show that an equilibrium is reached in which

$$T \approx T_0(1+\delta)^{\gamma-1}, \tag{4.25}$$

where T_0 is a normalization constant, and $\gamma \approx 1.6$ long after reionization. Because the photoheating rate is independent of density, this slope depends only on the varying dynamics of the expansion rate and so can be predicted robustly using linear theory [$\delta \propto D(t)$]. The normalization of this *temperature–density relation*[v], however, depends on the amplitude of that photoheating rate, which (by equation 4.21) is entirely determined by the *spectral shape* of the ionizing background. Importantly, this normalization is independent of the amplitude of the radiation background, because the heating rate per neutral atom is proportional to J_ν, but the neutral fraction is itself proportional to $1/J_\nu$ (through Γ

[iv] We will see later, however, that there is *on average* a nontrivial temperature–density relation during reionization, because the cosmic time at which elements are ionized depends on the density; see §9.9.

[v] This relation is sometimes referred to as the "IGM equation of state," but that is a misnomer because the relation implicitly averages over many different gas parcels, rather than following a single one.

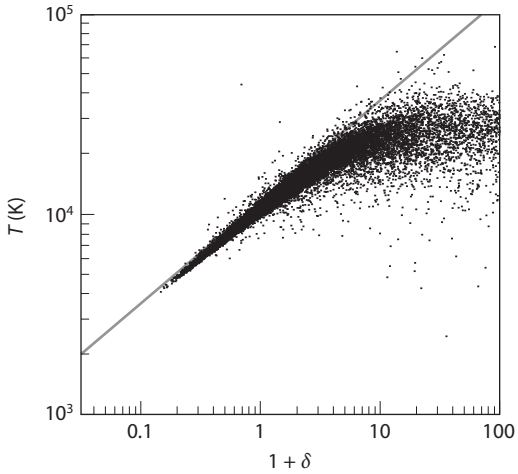


Figure 4.7 Temperature–density relation at $z = 4$ in a cosmological simulation (points), together with analytic prediction based on linear perturbation theory and equation (4.22). Note the near-power-law relation, except at very high densities, where the complex physics of nonlinear structure formation begins to matter. Hui, L., & Gnedin, N. Y. *Mon. Not. R. Astron. Soc.* **292**, 27 (1997). Copyright 1997 by the Royal Astronomical Society.

in equation 4.20). Figure 4.7 shows an example of this temperature–density relation. The points show individual cells from a cosmological simulation, while the solid line shows an analytic prediction using linear perturbation theory and equation (4.22), which leads to a solution of the form in equation (4.25). Note the very tight relationship at low and moderate densities in the IGM. The flattening at high densities is due to shock formation in filament and halo collapse, which is not included in our model here.

4.3.2 The Fluctuating Gunn-Peterson Approximation

A simple model for the absorption pattern of the inhomogeneous IGM associates each gas element with its "local" Gunn-Peterson optical depth in equation (4.11). This is an oversimplification for two reasons: first, it ignores the frequency structure of the line (in reality, the total τ_α at a given wavelength arises from many neighboring gas elements), and second, it ignores the velocity structure of the IGM, which moves gas elements around in frequency space. Nevertheless, the model provides a simple description and a reasonable approximation to the parameter dependencies of the real Lyman-α forest.

With the assumption of ionization equilibrium (equation 4.20) and if we ~~us~~ing the approximate power-law form of the temperature–density relation

use / PE

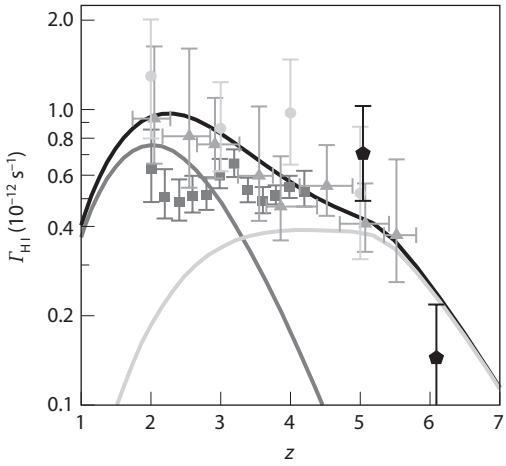


Figure 4.8 Measurements of the ionizing background at moderate redshifts (points with error bars), and models of it (curves). The measurements use the mean transmission of the Lyman-α forest as described in this section as well as the proximity effect (see §11.3.2). The curves base the emissivity and mean free path inputs on measured values at $z < 6$ and use the method described in §4.4. The solid line shows the total Γ, while the dashed curves peaking at $z \sim 4$ and $z \sim 2$ show the separate contributions from star-forming galaxies and quasars, respectively. Haardt, F., & Madau, P., *Astrophys. J.* **746**, 125 (2012) observational measurements collected from references therein. Reproduced with permission from the American Astronomical Society.

by / of / AAS. /

(equation 4.25), equation (4.11) becomes

$$\tau_\alpha(\delta, T) \approx 13 \frac{(1+\delta)^{2-0.76(\gamma-1)}}{\Gamma_{12}} \left(\frac{T_0}{10^4\ \mathrm{K}}\right)^{-0.76} \left(\frac{1+z}{7}\right)^{9/2}. \tag{4.26}$$

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

Equation (4.26) shows that at $z \sim 6$ only the most underdense regions will be visible (with $\tau_\alpha < 1$); gas at the mean density will be extremely opaque even if the ionizing background is comparable to its values at lower redshifts. This explains the deep absorption troughs in Figure 4.6. However, at $z \sim 3$ the same gas parcel at the mean density has $\tau_\alpha \sim 1$: this is why the Lyman-α forest is such a powerful tool at moderate and low redshifts, as shown in Figure 4.5.

Because τ_α depends only on fundamental cosmological parameters (which are known reasonably well), the density and temperature of the IGM (which can be modeled), and the unknown Γ, the transmission in the Lyman-α forest provides a good measure of Γ. The points with error bars in Figure 4.8 show several measurements of the ionizing background from $z \sim 2$–6, most using this method. To a reasonable approximation $\Gamma_{12} \sim 1$ over the range $z \sim 2$–5, with uncertainties (both systematic and statistical) of a factor ~ 2. The solid

curves show a theoretical prediction of $\Gamma(z)$, calculated following the methods we will outline in the next several sections.

Equation (4.26) shows that the IGM absorption traces the density field of the IGM, and hence the cosmic web. Because of this inhomogeneous absorption, the mean transmission averaged over broad spectral bands is *smaller* than through a homogeneous one. To see this difference, let us define $p_v(\delta)$ as the volume-averaged probability distribution of the IGM density. Then, the net transmission is

Delete subscript v x2 / PE

$$T_\alpha = \int d\delta \; p_v(\delta) \exp[-\tau_\alpha(\delta)] \tag{4.27}$$

$$\equiv \exp(-\tau_{\mathrm{eff},\alpha}), \tag{4.28}$$

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

$$\left\langle \exp(-\tau_\alpha) \right\rangle \geq \exp(-\langle \tau_\alpha \rangle). \tag{4.29}$$

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

4.3.3 The Column Density Distribution

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

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

$$N_{\mathrm{H\,I}} = 3.3 \times 10^{14}\,\mathrm{cm}^{-2} (1+\delta)^{3/2} \left(\frac{T_0}{10^4\,\mathrm{K}}\right)^{-0.26} \Gamma_{12}^{-1} \left(\frac{1+z}{7}\right)^{9/2}. \tag{4.30}$$

As described previously, the properties of these regions are measured from their optical depth for Lyman-α absorption, which for a single absorbing system is $\tau_\alpha(v) = N_{\mathrm{H\,I}} \sigma_\alpha(v)$. To understand the density distribution of the IGM,

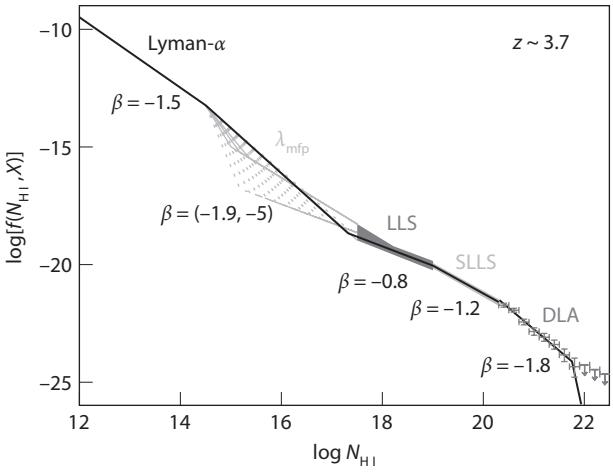


Figure 4.9 Observational estimate of the Lyman-α forest column density distribution function $f(N_{\rm H\,I})$ at $z \approx 3.7$, fit as a series of broken power laws. The different column densities are labeled according to their categories: Lyman-α for optically thin forest systems, LLS for Lyman-limit systems, SLLS for "super Lyman-limit systems", and DLA for damped Lyman-α systems. The distribution in the remaining region, labeled $\lambda_{\rm mfp}$ is difficult to probe directly and so is inferred indirectly. That region therefore has relatively large error bars compared with regions of higher column densities (which are shown by the shaded regions or points with error bars). Prochaska, J. X., O'Meara, J. M., & Worseck, G., *Astrophys. J.* **718**, 392 (2010). Reproduced with permission of the American Astronomical Society.

Should be hyphen, no spaces / PE

by /

AAS. /

we therefore would like to measure the the number density of absorbers in a column density interval $(N_{\rm H\,I}, N_{\rm H\,I} + dN_{\rm H\,I})$ and in a redshift interval $(z, z + dz)$, $d^2N/dN_{\rm H\,I}\,dz$. In the literature, the column density distribution is often reported as $f(N_{\rm H\,I}, z) = d^2N_{\rm H\,I}/dn_{\rm H\,I}\,dX$, where the coordinate X is defined via the differential relation $dX/dz = H_0(1+z)^2/H(z)$. This coordinate is useful because a population with constant comoving number density and constant proper cross section will have f independent of redshift. In the past, $d^2N/dN_{\rm H\,I}\,dz \propto N_{\rm H\,I}^{-\beta}$, where $\beta \approx 1.5$, was often used as a convenient and simple fit to the forest data. However, Figure 4.9 shows a recent measurement of this function at $z \sim 3.7$, which indicates that a single power law may be too simple an approximation, though a broken power law does fit quite well.

g / PE

Typically, one estimates this distribution function by identifying each absorbing system and fitting its column density. However, in practice, $N_{\rm H\,I}$ can be difficult to measure because of saturation. The *equivalent width* parameterizes the amount of absorption by specifying the wavelength interval over which light would be absent if the line profile were a step function,

$$W = \int [1 - e^{-\tau(\lambda)}]\,d\lambda. \tag{4.31}$$

When τ is small, $W \propto \int d\lambda \tau(\lambda) \propto N_{\rm H\,I}$, which makes the measurement straightforward. This width corresponds to column densities $N_{\rm H\,I} < 10^{14.5}$ cm^{-2}, or $\tau_\alpha < 1$ at line center. These systems are known as *Lyman-α forest absorbers*, although the forest is also often taken to mean the full set of absorbers (as we have used the term earlier in this chapter). They appear at the upper left corner of Figure 4.9.

However, when $\tau_\alpha \gg 1$, the line center is strongly saturated but the Lorentzian wings (see equation 4.8) have substantial optical depth and dominate the total absorption; in that case, $W \propto N_{\rm H\,I}^{1/2}$. In particular, absorbers with extremely high column densities ($N_{\rm H\,I} > 10^{20.3}$ cm^{-2}) have prominent damping wings from natural line broadening and are known as *damped Lyman-α absorbers* (DLAs); they appear at the lower right in Figure 4.9, and an example absorber is shown in the top panel of Figure 4.10. These systems, although rare, are extraordinarily rich in information. They have multiple absorption components (at slightly different velocities), a wide range of metal lines (with a wide range of ionization states), and sometimes even molecular hydrogen.

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

At somewhat lower column densities, the damping wings are less apparent, but the column density can indirectly estimated be through its effect on ionizing photons. At the ionization threshold, $\tau_{\rm H\,I} = 1$ for $N_{\rm H\,I,LLS} \equiv 1/\sigma_{\rm H\,I}(\nu_{\rm H\,I}) = 1.6 \times 10^{17}$ cm^{-2}. Systems above this column density limit are opaque to ionizing photons; we refer to this regime as *self-shielding* and these opaque systems as *Lyman-limit systems* (LLSs). The former term suggests that gas on the outskirts of the system absorbs a large fraction of the incident ionizing background, shielding the interior from ionizing photons.

For this reason, these systems are relatively easy to identify even at high redshifts, because their optical thickness to ionizing photons causes a continuum depression in the flux of the background source blueward of 912 Å in the rest frame of the absorber. An example is shown in the right panel of Figure 4.5; the sharp break at this wavelength indicates the presence of an LLS, and the depth of the absorption tells us the column density of the absorber. Because of this continuum suppression at very short wavelengths, these systems constitute the one family of hydrogen absorbers whose abundance at $z > 5$ has been measured. These, and their somewhat higher column cousins called

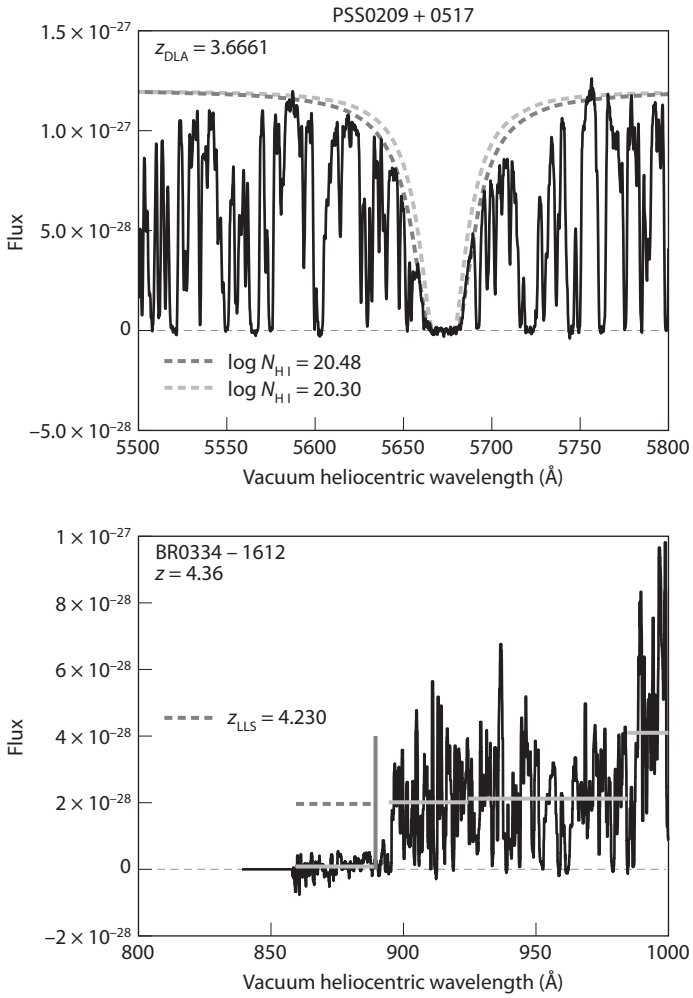


Figure 4.10 Example IGM Lyman-α absorbers measured by the ESI spectrograph on the Keck telescope. *Top:* A damped Lyman-α absorber (DLA). The two curves show fits to the absorber; note the long damping wings. *Bottom:* A Lyman-limit system (LLS) that is opaque to ionizing radiation (with $\lambda < 912$ Å). The horizontal bars redward of 912 Å show fits to the quasar continuum; the dashed line blueward of this wavelength shows the expected flux. The observed flux is much smaller because of the LLS, whose location is marked with the vertical line. Songaila, A., & Cowie, L., *Astrophys. J.* **721**, 1448 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

super Lyman-limit systems, appear from $10^{17.8}$ cm$^{-2} < N_{\mathrm{H\,I,LLS}} < 10^{20.3}$ cm^{-2} in Figure 4.9.

Unfortunately, in the intermediate regime where the line center is saturated but the wings remain weak, $W \propto \ln N_{\mathrm{H\,I}}$. When the opacity to ionizing photons

is also small, it is very difficult to measure the true column density of a line.[vi] This intermediate range approximately spans H I column densities of $10^{14.5}$–$10^{17.5}$ cm^{-2}. Their abundance must usually be inferred indirectly; Figure 4.9 shows the allowed range of density distributions at $z \sim 3.7$ by the series of dashed lines.

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

$$\tau_{\rm eff,\alpha} = \frac{(1+z)}{\lambda_\alpha} \int dN_{\rm H\,I} \frac{d^2N}{dN_{\rm H\,I}\,dz} W(N_{\rm H\,I}). \quad (4.32)$$

Note that $\tau_{\rm eff,\alpha}$ is *not* simply the average of the optical depths of all the lines, because the observed transmission depends exponentially on τ_α. Importantly, however, $\tau_{\rm eff,\alpha}$ does not *require* a measurement of $d^2N/dN_{\rm H\,I}\,dz$: as the total absorption, it can be estimated even from low-resolution measurements or when the forest is so thick that Lyman-α absorbers cannot be separated. This makes it a very useful quantity at high redshifts, where the forest is nearly saturated (see §4.7).

Observations show that at $z < 5.5$[8]

$$\tau_{\rm eff,\alpha} = (0.85 \pm 0.06) \left(\frac{1+z}{5}\right)^{4.3\pm0.3}. \quad (4.33)$$

At $z > 5.5$ the optical depth appears to increase even more rapidly and is nearly saturated at $z \sim 6$; we discuss this regime in §4.7.

4.3.4 Mapping the Cosmic Web

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

Thus, an important caveat for Lyman-α forest studies of the high-z Universe is that although detailed structures are visible along the line of sight, inverting these to obtain the three-dimensional structure is difficult because of *aliasing* (see Figure 4.11). This refers to the possibility that random arrangements of small-scale oscillations inclined to the line of sight will mimic large-scale oscillations along the line of sight; for example, if the crests of two k-modes are aligned with the plane of the sky (but at a wide radial separation) and intersect the Lyman-α forest skewer, they will appear to an observer as two crests of a single large-wavelength oscillation along the line of sight.

To quantify the importance of aliasing, we begin with the correlation function: statistical isotropy guarantees that it is identical in every direction and

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

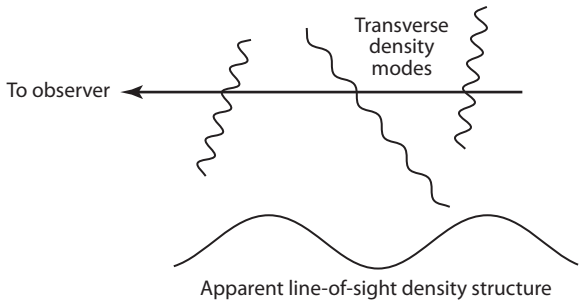


Figure 4.11 Illustration of *aliasing* in the Lyman-α forest. The observer can measure structure only along the line of sight. However, density modes transverse to that skewer can still leave imprints in the one-dimensional field where they intersect it. The superposition of many such modes can mimic large-scale fluctuations along the line of sight, *aliasing* power from small to large scales.

2.19

Roman / AA

so can be measured with data along only the line of sight. This function is related to the three-dimensional power spectrum P_{3D} through a Fourier transform (equation ~~2.20~~). However, if we use the Lyman-α forest data alone to measure a power spectrum, we obtain only a one-dimensional power spectrum, P_{1D}. This is *not* the same as P_{3D}, as the following argument illustrates. Let $k_\parallel$ and x be the wavenumber and distance coordinate along the line of sight. Then,

$$P_{1D}(k_1) = \int dx\, \xi(x) e^{ik_1 x} \tag{4.34}$$

$$= \int dx\, e^{ik_1 x} \int \frac{d^3k}{(2\pi)^3} P_{3D}(k) e^{-ikx}. \tag{4.35}$$

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

$$P_{1D}(k_1) = \int \frac{dk_y\, dk_z}{(2\pi)^2} P_{3D}(\sqrt{k_1^2 + k_y^2 + k_z^2}) \tag{4.36}$$

$$= \int_{|k_1|}^{\infty} \frac{dk_\perp}{2\pi} k_\perp P_{3D}(k_\perp), \tag{4.37}$$

Insert comma before theta / PE

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

Thus, the Lyman-α forest is best at constraining cosmological information on small physical scales. Of course, it is precisely these scales that are most difficult to model, so numerical simulations are necessary for quantitative constraints on, for example, the matter power spectrum. This procedure also helps constrain astrophysical parameters that affect the forest—most importantly, the ionizing background (which sets the overall normalization) and the

temperature (which sets the maximum wavenumber of interest through thermal broadening and Jeans smoothing of the IGM features). Aliasing is particularly important at high redshifts, where the forest is saturated, so that small fluctuations are magnified in importance.

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

4.4 The Metagalactic Ionizing Background

If we presume that one can model the structure of the IGM reliably, the primary physical input determining the opacity of the Lyman-α forest is the ionization rate Γ, which in turn depends on the angle-averaged specific intensity of the radiation background, $J(\nu)$ (equation 4.13).

The equation of cosmological radiative transfer determining the evolution of $J(\nu)$ can be derived from simple arguments. First, consider the total number of photons in an infinitesimal volume ΔV and frequency range $\Delta\nu$ at a time t, $N = n_\nu \Delta V \Delta\nu$, where n_ν is the photon number density per unit frequency ν. At some later time $t + \Delta t$, this total number N must be conserved provided photons are not absorbed or created, although both the photon frequencies and volume will have changed by the cosmic expansion. For example, the frequency interval becomes

$$\Delta\nu(t + \Delta t) \approx \Delta\nu(t) + \Delta t \frac{d\Delta\nu}{dt} \tag{4.38}$$

$$\approx \Delta\nu(1 - H\Delta t). \tag{4.39}$$

Performing a similar operation on the volume factor and using the constancy of N, we find

$$\frac{dn_\nu}{dt} = -2Hn_\nu. \tag{4.40}$$

Noting that n_ν is a function of both frequency and time, we therefore have

$$\frac{\partial n_\nu}{\partial t} = H\nu\frac{\partial n_\nu}{\partial \nu} - 2Hn_\nu. \tag{4.41}$$

Finally, transforming to the specific intensity $J_\nu = (c/4\pi)h\nu n_\nu$, and allowing for absorption and emission along the ray (just as in the standard radiative transfer equation), we find

$$\frac{\partial J_\nu}{\partial t} - \nu H\frac{\partial J_\nu}{\partial \nu} + 3HJ_\nu = -c\kappa_\nu J_\nu + \frac{c}{4\pi}\epsilon_\nu, \tag{4.42}$$

where κ_ν is the absorption coefficient, and $\epsilon(\nu)$ is the proper emissivity (in units of erg s^{-1} cm^{-3}). Here, the second term on the left-hand side accounts for the cosmological redshift, and the third for the dilution of the photons due to the

subscript greek nu / AA

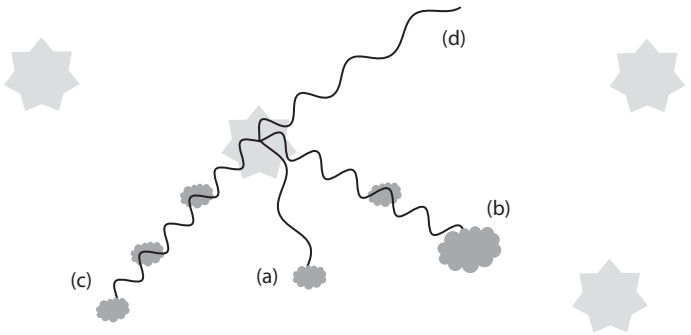


Figure 4.12 Radiative transfer of ionizing photons through the IGM. Here, a central source emits photons through the clumpy IGM. Photons near the ionization threshold (*a*) are easily absorbed even by low–column density systems, called *Lyman-limit systems*. These systems are typically transparent to higher-energy photons (*b*), which require higher column densities for a substantial optical depth. Nevertheless, these photons can still be absorbed by the accumulated opacity of low–column density systems (*c*). If even these clouds are rare, photons will simply redshift until they pass below the ionization threshold (*d*). Calculating these opacities is difficult because they depend on the background from many nearby quasars—which in turn depends on the mean free path of the photons, and hence the ionization state of the clouds.

cosmic expansion, while the first term on the right-hand side describes absorption as the photon passes through the IGM gas.

Integration of this equation gives the intensity at an observed wavelength ν and redshift z:

$$J_\nu(z) = \frac{c}{4\pi}\int_z^\infty dz' \left|\frac{dt}{dz'}\right| \left(\frac{1+z}{1+z'}\right)^3 \epsilon_{\nu'}(z') e^{-\tau_{\rm eff,H\ I}(\nu',z')}, \qquad (4.43)$$

where $\nu' = \nu(1+z')/(1+z)$ is the frequency of the photon at z' that will have frequency ν at z, and $\tau_{\rm eff,H\ I} \equiv -\ln\langle\exp(-\tau_{\rm H\ I})\rangle$ is the effective optical depth of an ionizing photon as it travels through the Universe (defined similarly as in equation 4.28). As usual, here we have ignored helium, which should be included in the optical depth calculation and strongly affects high frequencies (see §4.5). This optical depth depends on frequency and on the ionization state of the IGM, as depicted in Figure 4.12. If the mean free path of an ionizing photon is very short, so that cosmological effects can be neglected, and if we assume that $\tau_{\rm eff,H\ I} = r/\lambda(\nu, z)$, so that it scales linearly with distance traveled, this equation simplifies to

$$J(\nu, z) = \int_0^\infty 4\pi r^2 dr \frac{\epsilon(\nu, z)}{(4\pi r)^2} e^{-r/\lambda(\nu,z)} = \frac{1}{4\pi}\epsilon(\nu, z)\lambda(\nu, z), \qquad (4.44)$$

which shows that the two key inputs for this calculation are simply a measure of the absorption and the emissivity. This is a reasonable approximation at high redshifts, except for the highest-energy photons.

subscript greek nu x2 / AA

The emissivity clearly depends only on the sources—galaxies and quasars—and understanding this coefficient is a key goal of the following chapters. In brief, stellar sources typically have relatively soft spectra: hot stars with a surface temperature $\sim 30{,}000$ K, for example, have their blackbody peak at $E \sim 7$ eV, and their emission luminosity declines sharply at higher photon energies. The spectrum of solar-mass stars cuts off well before the Lyman limit and does not contribute significantly to the ionizing photon budget. Thus, because hot massive stars have such short lifetimes, only actively star-forming galaxies contribute to the metagalactic background. However, even their photons must escape absorption by the gas and dust inside their interstellar medium; this appears to be a difficult step in most known galaxy populations, where the so-called escape fraction is only a few percent and provides a key uncertainty in ab initio estimates of the ionizing background (see §9.2.2).

Quasars, the second important class of sources, are somewhat easier to model, partly because they are brighter and hence easier to observe. Their intense, high-energy radiation fields—typically with power-law spectra extending to X-ray energies—produce many more ionizing photons per unit bolometric energy output and probably allow a much larger fraction of these photons to escape to the IGM. In practice, both kinds of sources appear to be important at moderate redshifts (see Figure 4.8). However, beyond $z \sim 4$ the bright quasar population begins to decline precipitously, while the comoving star formation rate remains similar in magnitude. The natural expectation is therefore that galaxies become increasingly important at high redshifts.

4.4.1 The Mean Free Path of Ionizing Photons

The optical depth factor is determined by absorption in the IGM—and hence the Lyman-α forest. If we ignore line absorption processes, so that only bound-free opacity contributes to $\tau_{\mathrm{eff,H\ I}}$, we can easily write an expression for the mean free path from the column density distribution. The total opacity of the IGM per unit redshift at a frequency ν is just the sum of the opacities of all the individual absorbers,

$$\frac{d\tau_{\mathrm{eff,H\ I}}(\nu)}{dz} = \int dN_{\mathrm{H\ I}} \frac{d^2N}{dN_{\mathrm{H\ I}}\, dz}[1 - \exp \tau(\nu, N_{\mathrm{H\ I}})], \tag{4.45}$$

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

$$\lambda(\nu, z) = \frac{dr/dz}{d\tau_{\mathrm{H\ I,eff}}(\nu)/dz}, \tag{4.46}$$

where dr/dz is the comoving line element.

Given the distribution function of Lyman-α absorbers, this is a well-posed calculation, so it might appear to be straightforward to predict the mean free path from first principles. However, recall that τ_α is itself a function of Γ, which in turn depends on $\lambda(\nu, z)$. Self-consistently predicting the attenuation—and with

Insert { - / AA (left curly brace minus sign)

insert } / AA (right curly brace)

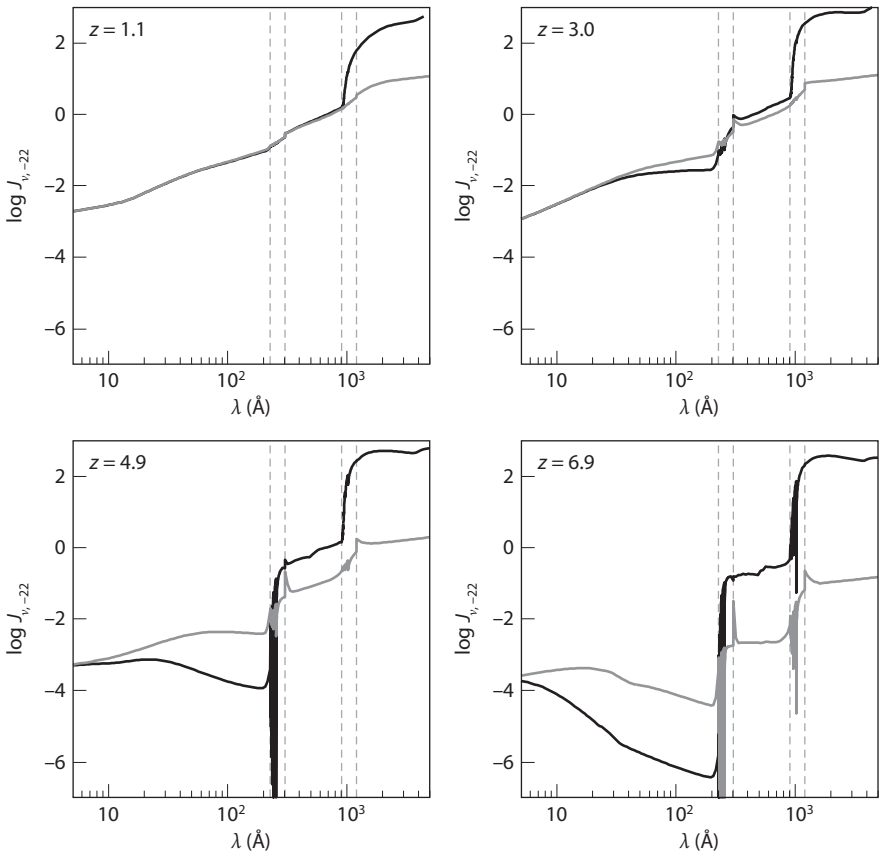


Figure 4.13 Metagalactic ionizing background spectrum at $z = 1.1$, 3.0, 4.9, and 6.9, in the same theoretical model as Figure 4.8. The units correspond to $J_{\nu,-22} = J_\nu/(10^{-22}$ erg cm^{-2} s^{-1} Hz^{-1} sr$^{-1})$. The dark and light lines show models with galaxies and quasars, and only quasars, respectively. The vertical lines bracket the Lyman-series transitions of H I and He II. Haardt, F., & Madau, P., *Astrophys. J.*, **746** 125 (2012). Reproduced with permission of the American Astronomical Society.

by / AAS. /

it the ionizing background—is therefore a rather complex problem that must be attacked iteratively. Figure 4.8 shows the resulting $\Gamma(z)$ from one such calculation. The solid curve shows the net ionizing background, while the two dashed curves show the contributions from quasars and star-forming galaxies (the former peaks at $z \sim 2$; the latter assumes an escape fraction that increases rapidly toward higher redshifts). Note that the total quasar emissivity is well measured to $z \sim 5$, and the rapid decline at high redshifts $z > 3$ is unavoidable. Thus, it appears that stars and quasars are both important for the ionizing background, each dominating at a different cosmological epoch.

Figure 4.13 shows the metagalactic ionizing background spectrum in the same calculation at four different redshifts. In each panel, the dark and light

curves show models with and without star-forming galaxies; including these dramatically softens the spectrum, especially at high redshifts. The light vertical lines demarcate the wavelength range between the Lyman-α line and ionization edge for H I and He II. This calculation includes line transfer effects, which are quite important in these regimes when the gas is not fully ionized (see §6.1.4). The sharp break at the H I ionization edge occurs because the Universe is so optically thick to these photons.

Does most of the opacity originate from these high–column density systems with $\tau_{\rm H\,I} > 1$ or from the accumulated opacity of lower–column density systems? Let us suppose for simplicity that $d^2N/dN_{\rm H\,I}\,dz = A[N_{\rm H\,I}\sigma_{\rm H\,I}(\nu)]^{-\beta} \approx A\tau_{\rm H\,I}^{-\beta}(\nu/\nu_{\rm H\,I})^{-3\beta}$, where $\tau_{\rm H\,I}$ is defined at the ionization threshold; as discussed previously, a single power law with $\beta \approx 3/2$ provides a crude but reasonable approximation to the observed distribution. The mean free path is then (for $1 \le \beta \le 2$)

$$\lambda(\nu) = \left[\frac{A}{\sigma_{\rm H\,I}(\nu_{\rm H\,I})}\right]^{-1}\left(\frac{\nu}{\nu_{\rm H\,I}}\right)^{-3(1-\beta)}\left[\int_0^\infty d\tau_{\rm H\,I}\,\tau_{\rm H\,I}^{-\beta}(1-e^{-\tau_{\rm H\,I}})\right]^{-1} \quad (4.47)$$

$$\approx \frac{1}{\Gamma_G(2-\beta)}\lambda_{\rm LLS}\left(\frac{\nu}{\nu_{\rm H\,I}}\right)^{-3(1-\beta)}. \quad (4.48)$$

Here, $\Gamma_G(x)$ is the Gamma function (not to be confused with the ionization rate). We have assumed that the absorbers span the range from $\tau_{\rm H\,I} \ll 1$ to $\tau_{\rm H\,I} \gg 1$ (with a single power law), and $\lambda_{\rm LLS}$ is the mean free path at the ionization edge including absorption only from systems with $\tau_{\rm H\,I} > 1$ (we normalize to this value because it is relatively easy to measure). For $\beta = 1.5$, $\lambda(\nu_{\rm H\,I}) \approx 0.56\lambda_{\rm LLS}$, so $\sim$56% of the absorption comes from the opaque systems. However, Figure 4.9 shows that at $z \sim 3.7$ the observed distribution flattens significantly throughout (at least) the upper end of the LLS regime, with $\beta \approx 1.2$. In that case, $\lambda(\nu_{\rm H\,I}) \approx 0.86\lambda_{\rm LLS}$, and the high–column density systems provide much more of the opacity. The evolution and distribution of these LLSs is therefore crucial to understanding the ionizing background.

Clearly, the mean free path is much longer for high-energy photons: with the canonical value $\beta = 3/2$ we have $\lambda \propto \nu^{3/2}$. This is a much weaker dependence than the $\lambda \propto \nu^3$ expected in a uniform IGM—its clumpiness is crucial in regulating the high-energy background. Indeed, only DLAs can efficiently absorb photons in the X-ray regime.

The redshift evolution of LLSs therefore provides important insight into the ionizing background, even if these systems may not dominate the total absorption. Fortunately, as we saw in §4.3.3, they are relatively easy to identify even at high redshifts. Recent surveys have established the LLS abundance reasonably well at $0 < z < 6$; the additional assumption that $d^2N/dN_{\rm H\,I}\,dz \propto N_{\rm H\,I}^{-\beta}$ with $\beta \approx 1.1$–1.5 (and constant with redshift) yields a mean free path at the Lyman edge of[9]

$$\lambda(\nu_{\rm H\,I}) \approx (50 \pm 10)\left(\frac{1+z}{4.5}\right)^{-4.44\pm0.3}\ \text{proper Mpc}. \quad (4.49)$$

Extrapolation to $z \sim 6$ (10) yields a mean free path of ~ 7 (1) proper Mpc—during the era of the first galaxies, ionizing photons suffered *much* more attenuation than at later times. This obviously led to strong fluctuations in the ionizing background and substantially affected the process of reionization. We will consider the importance of LLSs in more detail later: in essence, they regulated the end of reionization and provided the "matching condition" from the epoch of reionization to later times (see §9.5).

Despite the relative ease of *finding* LLSs, their physical nature remains obscure. Although the self-shielded absorbers are opaque to photons at the ionization edge, the strong frequency dependence of the ionization cross section implies that they are transparent to higher-energy photons. Typically, this implies that LLSs are themselves highly ionized. If we assume that a system with column density $N_{\rm H\ I}$ is opaque to all photons with $\nu < \nu_{\rm min}$ and that $\epsilon(\nu) \propto \nu^{-\alpha}$, we have an effective ionization rate reaching the interior of the absorbers $\Gamma \propto [N_{\rm H\ I}\sigma_{\rm H\ I}(\nu_{\rm H\ I})]^{(-\alpha+3\beta-6)/3}$, so according to equation (4.20) the residual neutral fraction is

$$x_{\rm H\ I} \sim 2.2 \times 10^{-4} T_4^{-0.59} \Gamma_{12}^{-1/3} [N_{\rm H\ I}\sigma_{\rm H\ I}(\nu_{\rm H\ I})]^{(-\alpha+3\beta-6)/3}. \quad (4.50)$$

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

$$1 + \delta_{\rm LLS} = 320\, T_4^{0.17} \Gamma_{12}^{2/3} \left(\frac{1+z}{4}\right)^{-3}, \quad (4.51)$$

which follows from equation (4.30) if we set $N_{\rm H\ I} = N_{\rm H\ I,LLS}$.

Equation (4.51) shows that at moderate redshifts these objects have overdensities comparable to those inside virialized halos. As such, they are difficult to model and require high-resolution numerical simulations of the structure of gas around galaxies, coupled with a large enough cosmic volume, to represent adequately the cosmic radiation field. Explanations for their origin range from low-mass dark matter halos without substantial star formation to the accretion of cold gas onto galactic halos from filaments in the cosmic web.

To complicate matters further, the very nature of these LLSs may evolve at higher redshifts. Even if we assume optimistically that the ionizing background remains constant, equation (4.51) shows that $\delta_{\rm LLS} \sim 20$ at $z \sim 10$. This illustrates how dangerous it can be to make inferences about the epoch of reionization from extrapolations of features at low or moderate redshifts.

4.4.2 Fluctuations in the Ionizing Background

Because ionizing photons can travel only finite distances and are generated by discrete sources, one naturally expects fluctuations in the amplitude (and possibly shape) of the ionizing background. In practice, these are very small at $z < 5$, because λ is relatively large (see §4.4.1), and the ionizing sources are relatively common (particularly galaxies, provided their escape fraction of ionizing photons is nonzero). But these fluctuations inevitably become important at higher redshifts (especially toward the epoch of reionization).

These fluctuations are sourced by both large-scale density fluctuations and stochastic variations in the number counts of the sources; the latter are most important when the number of sources within $\sim\lambda^3$ is small. The effects of the density field can be simply esimated by computing the variance of the source population over one attenuation length, $\bar{b}\sigma(R=\lambda)$, where $\bar{b}$ is the average bias of the sources. At $z\sim 3$, taking $\bar{b}\sim 3$ and $\lambda\sim 300$ Mpc, we obtain fractional fluctuations of $\sim 2\%$, which is indeed close to more precise numerical estimates and is mostly negligible. However, at $z\sim 6$, the same average bias but with $\lambda\sim 50$ Mpc implies fluctuations of $\sim 10\%$. At earlier times, or if the mean free path is even smaller, these fluctuations will only increase. This change in fluctuation amplitude could be important for regulating the temperature of the IGM, which in turn affects the accretion of baryons onto galaxies (see §9.9). More sophisticated models for the fluctuations can be constructed using the halo model (§3.6.1) by replacing the halo density profile with the radiation intensity profile around each source (see §12.3.1 for a similar application). Such calculations show fluctuations comparable in magnitude to our simple estimate.[10]

4.5 The Helium-Ionizing Background

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

The ionization cross section for He II follows the same form as in equation (4.14). Like $\sigma_{\rm H\ I}$, this cross section also scales as ν^{-3} near threshold. He II is also more difficult to keep ionized because it recombines faster than hydrogen; its case-B recombination coefficient is $\alpha_B = 1.53\times 10^{-12}$ cm^3 s^{-1} at $T = 20{,}000$ K. The recombination timescale for gas at the mean density therefore remains smaller than the age of the Universe down to $z < 3$. Thus, He II atoms may recombine many times over the age of the Universe.

Other than making these changes to the input parameters (and the trivial change of including electrons from both helium and hydrogen), we can compute the helium-ionizing background $\Gamma_{\rm He\ II}$ in the same way as we did for H I. Because of the lower ionizing background and faster recombination rate, He II is usually significantly more common than H I, despite the lower intrinsic abundance of helium. This relationship is parameterized by

$$\eta \equiv \frac{N_{\rm He\ II}}{N_{\rm H\ I}} \approx 1.77\frac{\Gamma}{\Gamma_{\rm He\ II}}, \tag{4.52}$$

where in the second equality we have assumed that the system is optically thin to ionizing photons for both species. Even after helium reionization, $\Gamma_{\rm He\ II}\sim 10^{-14}$ s^{-1}, so $\eta\sim 100$.

Because of the large ionization cross section and rapid recombination time, the Universe remains optically thick to He II-ionizing photons until relatively late (or, in other words, the mean free path of these photons is many times shorter than that of photons below the He II ionization threshold). As a result, there is typically a substantial break in the ionizing background at the He II ionization edge until that species is fully ionized at $z \sim 3$ (see Figure 4.13). Moreover, the He II–ionizing background has much stronger fluctuations than the H I–ionizing background, both because of the short attenuation length and because only rare quasars contribute to it.

For the most part, the properties of this high-energy background have little effect on the H I Lyman-α forest; however, the photoheating that occurs as the helium is ionized affects the hydrogen as well. The process is identical with that described for H I reionization in equation (4.23) (except with $n_{\rm H} \rightarrow n_{\rm He}$). Moreover, the hard spectra of quasars quite efficiently inject energy into the helium gas, so (despite the relative rarity of helium atoms in the IGM) the total temperature increase can be comparable to that during hydrogen ionization. Once helium is reionized at $z \sim 3$, any influence of hydrogen reionization on the gas is largely erased.

4.6 Metal-Line Systems

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

The typical abundance of heavy elements in the IGM is small—with a median value $\langle Z \rangle \sim 10^{-3} Z_\odot$—but the absorption is still substantial. If we make the simple assumption that the metals are uniformly distributed, we can repeat the fluctuating Gunn-Peterson approximation. From equation (4.10), $\tau \propto n_{X_i} f_{\rm osc} \lambda_i$, where n_{X_i} is the number density of the relevant species, $f_{\rm osc}$ is the oscillator strength of the transition, and λ_i is its rest wavelength. We then find that the optical depth of an IGM patch to a given transition is

$$\tau_{X_i} = 0.097 f_i (1+\delta) \left(\frac{X}{3.6 \times 10^{-7}}\right) \left(\frac{f_{\rm osc}}{0.191}\right) \left(\frac{\lambda_i}{1548\ \text{Å}}\right) \left(\frac{1+z}{7}\right)^{3/2}, \quad (4.53)$$

where f_i is the fraction of the element in the appropriate ionization state, and X is the abundance by number of the element relative to hydrogen. The fiducial choices correspond to the stronger line in the C IV λ1548, 1551 doublet with $Z = 10^{-3}\ Z_\odot$; Table 4.1 lists several other important transitions for low- and high-z work. As we will see later, the assumption of a constant metallicity

Table 4.1 Important IGM Metal-Line Transitions.

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

* Member of doublet

throughout the IGM is most certainly wrong, but it may be reasonable on the scale of a single absorbing system.

Clearly, the optical depth can be substantial, even in relatively low density gas, provided the gas is in the appropriate ionization state. This makes absorption from metals, not just H I, a useful probe of the IGM. In particular, the fluctuations in metal abundance trace the production and dispersal of heavy elements after the Big Bang (due to star formation), and their ionization states probe the metagalactic radiation background.

In the diffuse IGM at low and moderate redshifts, the best lines for these purposes are highly ionized states of the most common heavy elements, especially carbon, silicon, and oxygen. These make a particular interesting probe of the radiation background near the He II ionization edge. C III and Si IV have ionization potentials of 47.888 and 45.142 eV, respectively; these two species should therefore evolve similarly, *unless* higher-energy photons are able to further ionize one but not the other. In fact, C IV and Si V require 64.492 and 166.77 eV to be ionized. The latter energy is relatively large, but once He II is ionized to He III, the universe becomes transparent to photons that can ionize C IV, and such photons are still relatively common. We might therefore expect C IV and Si IV to be relatively abundant absorbers until He II reionization is complete at $z < 3$, after which the C IV abundance is substantially reduced.

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

$$(1 + z_{\rm min}) = (1 + z_s)\frac{\lambda_\alpha}{\lambda_i}. \tag{4.54}$$

Absorbers in this range produce isolated absorption features against the red continuum of the source. Figure 4.5 shows several examples: most of the

absorption features redward of Lyman-α in the rest frame of the source are IGM metal lines (although peculiar velocities could introduce a small number of Lyman-α absorption features just beyond the source redshift); note that they are typically weaker than the forest features but nevertheless are clearly visible. Doublet transitions are particularly interesting, because they make the species causing the absorption easy to identify, even in some cases without our knowing anything about the H I absorption. A particularly important such doublet is C IV λ1548, 1551, some examples of which are shown in the top panels of Figure 4.14. The doublets clearly stand out redward of Lyman-α.

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

In neutral gas, which we expect to dominate very early in the history of the Universe, the relevant ions are different. For example, C I has an ionization potential of 11.26 eV, so provided there is a source of UV photons—even if the H~~/~~I ionizing photons are attenuated—carbon atoms will preferentially turn into C II (which has an ionization potential of 24.383 eV and so cannot be ionized in gas that is optically thick to hydrogen-ionizing photons). Iron, another common heavy element, and silicon occupy their first ionization state for similar reasons.

H I / PE

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

$$O^0 + H^+ \leftrightarrow O^+ + H^0 \tag{4.55}$$

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

Note that all these transitions relevant to neutral gas are singlets, so they are more difficult to identify than the C IV and Si IV lines in an ionized gas. This means that the transitions must be identified in combination with each other (or H I), with the additional complication that the different elements may have different abundances. The bottom panel of Figure 4.14 shows such a system at $z = 5.88$. In this case, absorption is detected in C II, O I, and Si II lines, which provides solid evidence for a real absorber. Note that this system has no apparent absorption in the high ionization states C IV or Si IV. At lower redshifts, systems that are self-shielded enough from the metagalactic background to host substantially neutral gas are DLAs, which are associated with galaxies and so always host highly ionized gas as well.

At moderate redshifts, C IV absorbers are the most commonly studied (primarily because they are the easiest to find), with metal absorption visible in

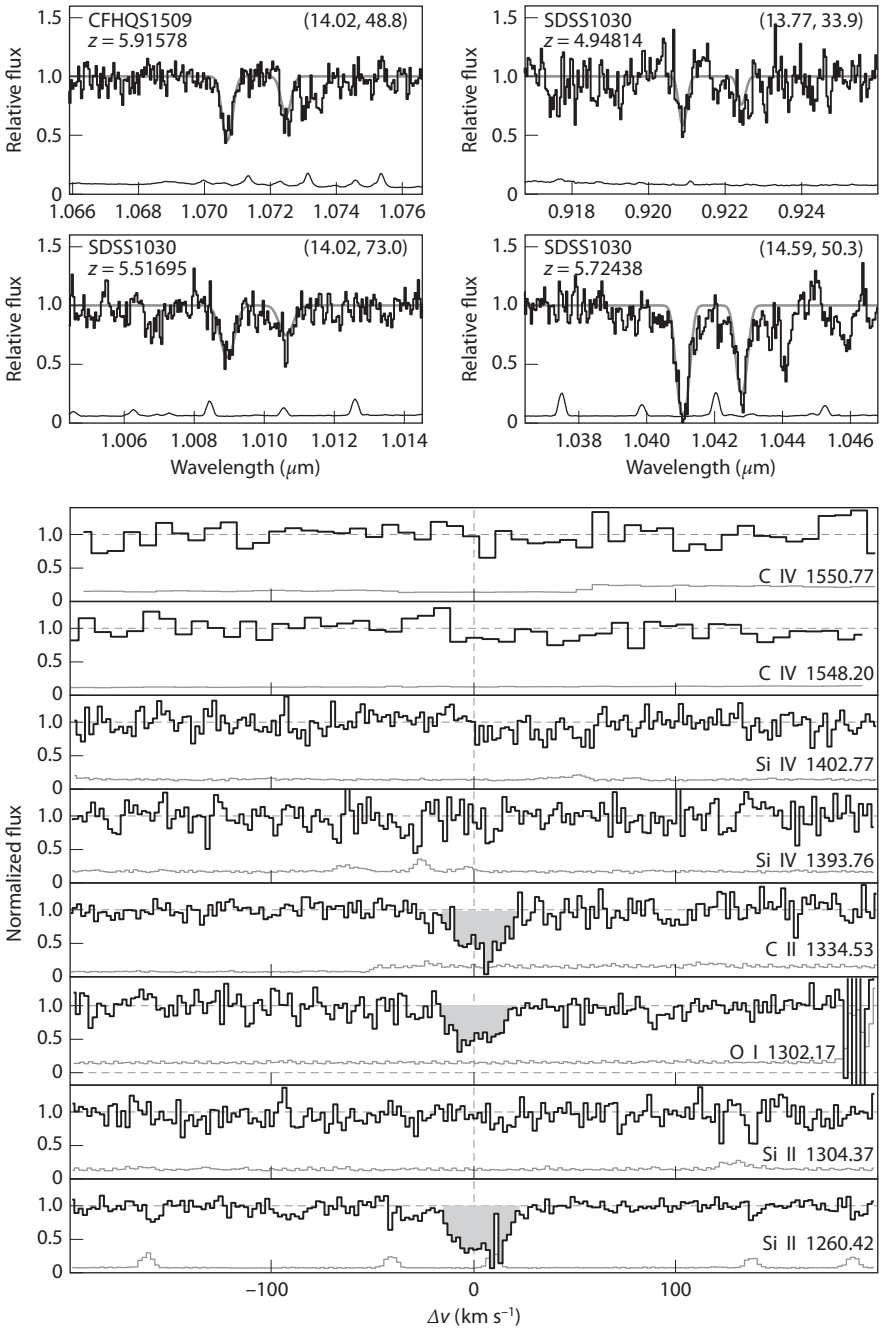


Figure 4.14 Examples of high-z metal absorption systems. *Top:* Several C IV doublets discovered with the Magellan Baade Telescope. The upper histograms show the data; the lower histograms show the errors. The solid curve in each panel is a Voigt profile fit to the two components of the doublet; the column

most individual systems with $N_{\rm H\ I} > 10^{15}$ cm^{-2} at a metallicity $Z \sim 10^{-2}\ Z_\odot$. They can also be detected statistically in much less dense systems, which implies a median metallicity in forest absorbers of $Z \sim 10^{-3}\ Z_\odot$.[11] Many other transitions are detectable in higher–column density systems, especially in the DLA range (where the neutral gas makes transitions like C II and O I useful, although these systems usually have many different absorption components, some of which are also highly ionized); these are well understood as being due to internal metal enrichment of galaxies. O VI has also received intense attention as a possible proxy for the hot, collisionally ionized gas in galactic winds.

Despite the relative wealth of observations of metal absorption, the physics behind metals in the IGM remains mysterious. The forest absorbers themselves correspond to gas near or above the mean cosmic density, and such sheets and filaments fill only a relatively small fraction of the volume. Thus, observations currently require only > 10% of space to be enriched with metals. The key question is how and when did this enrichment occur: many models appeal to winds from the first galaxies, but more powerful winds from star-forming galaxies at lower redshift are also a plausible explanation. More precise measurements of the spatial distribution of the metals (especially in comparison with samples of galaxies), their abundance patterns, and the evolution toward higher redshift may help constrain or eliminate some of these models, which we will discuss in §6.5.2.

4.7 The Lyman-α Forest at $z > 5$

We now turn to the Lyman-α forest at very high redshifts, approaching the time of reionization and the first galaxies. As equation (4.33) shows, the absorption is quite thick by $z \sim 5.5$ when $\tau_{{\rm eff},\alpha} \sim 2.6$, with only $\sim 7\%$ of the light transmitted. Past that point, the forest thickens even more rapidly, so that very little light is transmitted. Figure 4.15 shows this result in detail using measurements from quasars discovered with the Sloan Digital Sky Survey. Note the turn to higher effective optical depths at $z \sim 6$ in comparison with the smooth evolution at lower redshifts.

Of course, this low level of transmission is not uniform across the entire spectrum owing to the density fluctuations in the cosmic web. The small

Figure 4.14 (Continued.) density $N_{\rm C\ IV}$ and Doppler width b are listed in parentheses in each panel. *Bottom:* An example low-ionization absorber at $z = 5.8765$ toward the quasar SDSS J0818+1722 discovered with the Keck telescope. The solid dark lines are the data; the errors are shown as the lower lines in each panel. Shaded regions represent detected transitions. The H I lines are completely saturated in both absorbers. Simcoe, R. A., *et al. Astrophys. J.* **743**, 21 (2011, top panels); Becker, G. D., *Astrophys. J.* **735**, 93 (2011) astro-ph/1101.4399 (bottom panels). Reproduced ~~with~~ permission of the ~~American Astronomical Society~~.

by / AAS. /

COMP: Periods were used inconsistently. I have added missing periods as needed; these are PEs.

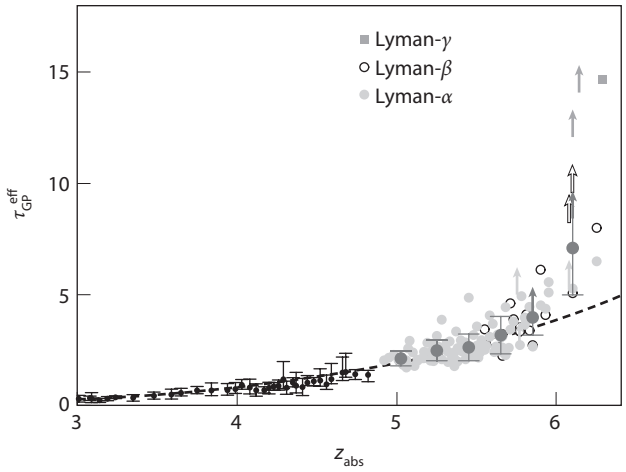


Figure 4.15 Measurements of the effective optical depth in the Lyman-α transition at high redshifts. The small filled circles at $z > 5$ are direct measurements from the Lyman-α forest, the open circles are measurements from Lyman-β absorption translated into Lyman-α, and the filled squares are the same for Lyman-γ. The large filled circles with error bars show the average inferred from the Lyman-α and Lyman-β measurements. At $z > 5.5$, $\tau_{\rm eff} \propto (1+z)^{11}$ or possibly even steeper. Fan, X., et al., *Astron. J.* **132**, 117 (2006). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

pockets of residual transmission correspond to underdense regions in the IGM. At $z \sim$ 5–6, these pockets are sufficiently common that the forest can still be used to measure the properties of the IGM, and in particular the ionizing background—which appears to be several times smaller than at lower redshifts (with $\Gamma_{12} \sim 0.2$; see Figure 4.8).

Unfortunately, beyond that point the Lyman-α forest itself becomes too thick to model robustly; in fact, it is so thick that one can no longer pick out individual absorbers, and it is more intuitive to use the fluctuating Gunn-Peterson approximation. If one then has a model for the volume-weighted probability distribution of the IGM density $p(\delta)$, the effective optical depth is simply given by equation (4.27). The function $p(\delta)$ is easy to describe qualitatively: it must peak near $\delta \sim 0$, with a long tail toward high densities (describing collapsed structures) and another tail toward underdense voids that is truncated below a value $\delta = -1$ (corresponding to space with no matter). Because all matter with $\Delta > 100$ is inside virialized objects, the density profile of these halos determines the shape of the high-δ tail: for power-law density profiles, equation (4.26) shows that with $\Gamma_{12} \sim 0.2$ at $z \sim 6$, $\tau_\alpha \sim 65(1+\delta)^2$ (if we ignore the weak temperature dependence), which requires $-1 < \delta < -0.8$ for measurable transmission. The leftmost solid line in Figure 4.16 illustrates this relationship explicitly, showing the contributions to the effective optical depth integral (equation 4.27), assuming a simple model for $p(\delta)$ and a uniform IGM

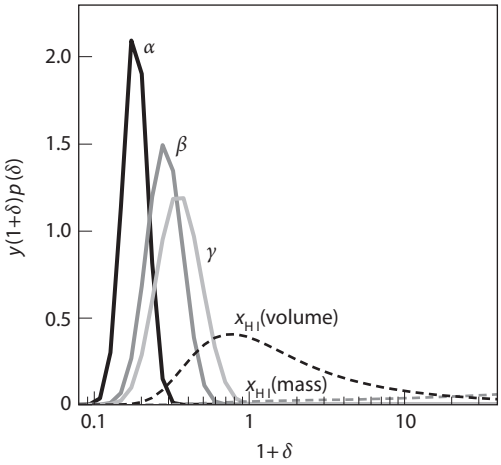


Figure 4.16 Logarithmic integrand $y(1+\delta)p(\delta) \propto \langle y \rangle$ in a simple model of the IGM density distribution. The three solid curves use the optical depth in the Lyman-α, Lyman-β, and Lyman-γ transitions (from left to right) for y; they thus show the integrands in equation (4.27) and its analogs for the higher-order transitions. The dashed lines show a similar calculation for the neutral fraction, with both volume and mass weighting (the latter peaks to the right of the plot). All curves assume a uniform $\Gamma_{12} = 0.04$ at $z = 6.15$ and an isothermal temperature–density relation. Oh, S. P., & Furlanetto, S. R., *Astrophys. J.* **620**, L9 (2005). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

temperature and ionizing background. Note that the transmission originates from very low density gas. In contrast, the dashed curves show the analogous integrand if one wished to compute the average neutral fraction in the same simple model: the overlap between these curves and the Lyman-α contribution is essentially the fraction of neutral gas sampled directly by the Lyman-α forest. Understanding the rest of the neutral gas requires extrapolation from this low-density regime.

Thus, the crucial piece of the integral involves the tip of the low-density tail (note that these voids are actually in the nonlinear regime), which is very difficult to model robustly without large numerical simulations. Even then, to measure the mean neutral fraction of the entire IGM one must extrapolate to significantly higher densities, which constitutes a highly uncertain operation. Conservatively, the observed transmission requires only a very small neutral fraction, $x_{H\,I} < 10^{-4}$, at the mean density.[12] Thus, the increasing optical depth of the forest with redshift is *not* necessarily a flag of the tail end of reionization; careful modeling of the forest is required to reach such a conclusion.

A few options can help improve this measurement and extend the usefulness of the Lyman-α forest to higher redshifts. The first is to use a different aspect of the forest: one probe that appears promising is to use large-scale variations in its optical depth, which may be modulated by the contrast between

neutral and ionized regions in the IGM. For example, some lines of sight at $z > 6$ show completely saturated absorption even in deep spectra, while others show clear transmission. Unfortunately, as described in §4.3.4, fluctuations in the absorption are dominated by the aliasing of small-scale modes in the density field, which tend to mask the underlying large-scale fluctuations. Moreover, the extremely underdense voids that allow transmission tend to lie in large-scale underdensities, which exaggerates their variance (i.e., they cluster just like rare, massive halos). Thus it is so far difficult to use these variations to constrain the neutral fraction quantitatively.

A second option is to use a higher Lyman-series line: so far, Lyman-β (with $\lambda_\beta = 1026$ Å) and Lyman-γ (with $\lambda_\gamma = 972$ Å) have been used. From equation (4.11), in a uniform medium $\tau \propto f_{\rm osc}\lambda_i$; for Lyman-$\beta$ and Lyman-γ, these factors give $\tau_\alpha/\tau_\beta = 6.25$ and $\tau_\gamma/\tau_\alpha = 17.9$—which on their face could provide a huge boost to the transmission. However, in practice, the observable $\tau_{\rm eff}$ changes much less, because the inhomogeneous IGM moderates the difference—in most models, $\tau_{\rm eff}$ increases only a few times.[13] However, Figure 4.16 shows that these transitions do require less extrapolation to moderate densities.

Of course, this improvement has a price. The primary difficulty is that the higher transitions are visible only at $\lambda_{\rm obs} < \lambda_{\beta,\gamma}(1 + z_s)$, which is inside the Lyman-α forest of the same source (albeit at a lower redshift, $z < \lambda_{\beta,\gamma}/\lambda_\alpha(1 + z_s)$, where the transmission is larger). One must therefore account for this unknown foreground absorption, which introduces extra errors. Nevertheless, the higher Lyman-series lines do appear to be more sensitive than Lyman-α, and they indicate a steepening in the effective absorption of the IGM and hence possible stronger evidence for an increasing neutral fraction at $z > 6$. Figure 4.15 shows this explicitly—the open circles and filled squares represent inferences from Lyman-β and Lyman-γ, respectively (converted to the expected Lyman-α optical depth using a simple model for the IGM). The measurements provided by these transitions are significantly more sensitive than Lyman-α on its own.

One complication regarding these lines is that because they probe slightly different densities than Lyman-α, they may also sample different temperatures if the gas is no longer isothermal ($\gamma \neq 1$). Indeed, as discussed in §4.3.1, the IGM is expected to have such a density–temperature relation once the gas relaxes after being heated during reionization. The temperature effects on the optical depth also makes inferences about Γ more difficult (see equation 4.26). However, it also offers a route to *measuring* this temperature–density relation and through it constraining the time of reionization (with the complication that denser regions may have reionized earlier than underdense regions; see §9.9.1).

Finally, instead of choosing weaker Lyman-series lines one can study rarer elements—the metal lines. With the forest saturated, it is no longer possible to associate these lines with H I features; however, they can still be detected individually as long as they appear redward of $\lambda_\alpha(1 + z_s)$ (see equation 4.54). Of course, one must then determine which species causes the observed line, for example, by detecting multiple absorbers from the same redshift. Although this wavelength range stretches into near-IR wavelengths for $z > 6$, such surveys are

possible with modern instruments and present an exciting frontier for pushing into the cosmic dawn.

Two other probes of the ionization state of the IGM are useful as more direct measurements of the reionization process: the so-called red damping wing (which refers to the Lyman-α absorption profile far to the red of line center, where the optical depth is of the order of unity even in a completely neutral medium) and the *proximity effect* (which refers to the highly ionized zone surrounding individual bright sources). We discuss these probes in chapter 11.

PART II

The First Structures

Chapter Five

The First Stars

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

The next section of this book considers the emergence of this complexity during the first stages of galaxy formation. We will study the appearance of the first stars, their feedback processes, and the resulting ionization structures that emerged during and shortly after the cosmic dawn. We start with a brief outline of the prevailing (though observationally untested) theory for this cosmological phase transition and then flesh out its details over the next two chapters.

As we saw seen in chapters 3 and 4, the development of large-scale cosmic structures occurs in three stages, as originally recognized by the Soviet physicist Yakov Zel'dovich. First, a region collapses along one axis, making a two-dimensional sheet. Then, the sheet collapses along the second axis, making a one-dimensional filament. Finally, the filament collapses along the third axis into a virialized halo. A snapshot of the distribution of dark matter at a given cosmic time should show a mix of these geometries in different regions that reached different evolutionary stages (owing to their different densities). The sheets define the boundary of voids whence their material was assembled; the intersections of sheets define filaments, and the intersections of filaments define halos—into which the material ultimately drains. The resulting network of structures, shown in Figure 4.1, delineates the so-called cosmic web. Gas tends to follow the dark matter (except within shallow potential wells, owing to its finite pressure).

Computer simulations have provided highly accurate maps of how the dark matter is expected to be distributed, since its dynamics is dictated only by gravity, but, unfortunately, this matter is invisible. As soon as ordinary matter is added, complexity arises because of its cooling, chemistry, and fragmentation into stars and black holes. Although theorists have a difficult time modeling the dynamics of visible matter reliably, observers can monitor its distribution through telescopes. The art of cosmological studies of galaxies involves a

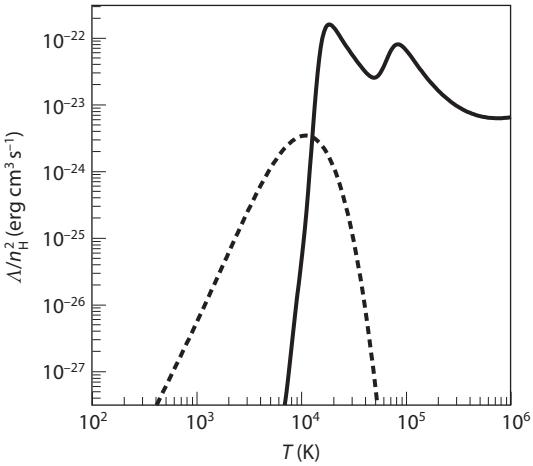


Figure 5.1 Cooling rates as a function of temperature for a primordial gas composed of atomic hydrogen and helium, as well as molecular hydrogen, in the absence of any external radiation. We assume a hydrogen number density $n_H = 0.045$ cm^{-3} corresponding to the mean density of virialized halos at $z = 10$. The plotted quantity Λ/n_H^2, where Λ is the volume cooling rate (in erg cm^3 s^{-1}), is roughly independent of density (unless $n_H > 10$ 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. Barkana, R., & Loeb, A., *Phys. Rep.* **349**, 125 (2001). Copyright 2001 by Elsevier.

delicate dance between what we observe but do not fully understand and what we fully understand but cannot observe. The next several chapters describe this methodology.

When a dark matter halo collapses, the associated gas falls in at a speed comparable to V_c in equation (3.31). When multiple gas streams collide and settle to a static configuration, the gas shocks to the virial temperature T_{vir} in equation (3.32)—at which it is supported against gravity by its thermal pressure. At this temperature, the Jeans mass equals the total mass of the galaxy. 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 *interstellar medium*, or ISM) is controlled by abundant heavy elements, such as carbon, oxygen, and nitrogen, which were produced in the interiors of stars. However, before the first stars formed no such heavy elements were present 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 had a virial temperature well below 10^4 K, 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 Kelvins (similar to the present temperature range on Earth). In this chapter, we focus on how this cooling took place and the properties of the stars that formed out of primordial gas.

However, this is far from the entire story, because once the first stars formed the initial conditions for *other* stars immediately became more complex. The feedback processes that set these conditions are the main subject of the next chapter. In particular, the hydrogen molecule is fragile and can easily be broken by UV photons (with energies in the range 11.26–13.6 eV), to which the cosmic gas is transparent even before it is ionized. 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 $T_{\rm vir}$.

To understand how structures proceeded from the first stars to subsequent generations, we must therefore understand *feedback processes*—in this case, UV and X-ray radiative feedback. We therefore examine in some detail the growth of these radiation backgrounds (in chapter 6) and how they may have affected star and galaxy formation. In this chapter, we also discuss how the chemistry of cooling changed dramatically when halos with $T_{\rm vir} > 10^4$ K formed. In such objects, atomic hydrogen was able to cool the gas and allow fragmentation even in the absence of H_2—such halos were thus immune to the radiation background.

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

Of course, because these same stars also produced heavy elements, which affected the chemistry and cooling of the gas, we must also track chemical feedback: how these elements were generated inside dark matter halos and how mechanical processes, most likely from supernovae or AGN, distributed these heavy elements within their parent halos and throughout the IGM (and hence the halos that assembled from it). We also discuss this enrichment process in chapter 6.

When these feedback mechanisms are included, the first structures to have formed stars likely could not continue to do so, at least for a time. Only later, possibly when atomic cooling became possible, would larger halos develop in

which self-sustaining "galaxies" could form. These long-lived objects will be much easier to observe than their predecessors and hence will provide an important marker in structure formation, especially for observers.

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

5.1 From Virialized Halos to Protostars

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

Specifically, in this section we consider stars forming out of primordial gas inside low-mass dark matter halos (in which the virial shock is not sufficient to ionize the infalling gas) without any influence from other stars or black holes in the Universe. These initial conditions led to what are now called *Population III.1 stars*; we consider an alternative pathway for primordial star formation in the next section. The basic steps described here are illustrated in Figure 5.2.

5.1.1 Chemistry of the Primordial Gas

The primordial gas out of which the first stars were made was 76% by mass hydrogen and 24% helium. It lacked any elements heavier than lithium, because the cosmic expansion rate was too fast to allow the synthesis of heavier elements through fusion reactions during Big Bang nucleosynthesis.

Before elements heavier than helium were produced in stellar interiors, the primary species to reach sufficient abundance to affect the thermal state of the pristine cosmic gas was molecular hydrogen, H_2.[1] The dominant H_2 formation process was

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

Italics / PE

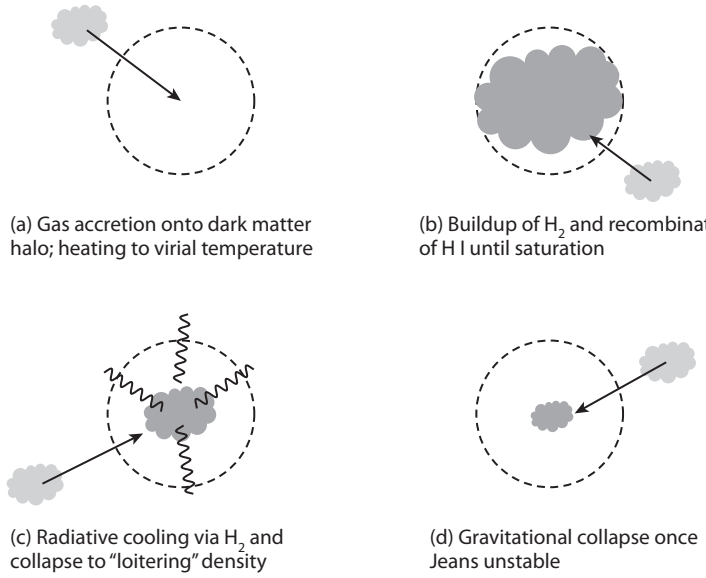


Figure 5.2 Basic stages in Population III.1 protostar formation. *(a)* A dark matter halo passes the filter mass threshold and begins to accumulate baryons. The virial shock (dashed line) heats the gas to the halo's virial temperature. *(b)* Inside the warm, dense gas, free electrons catalyze H_2 formation until those electrons disappear owing to recombinations, when the H_2 fraction saturates. *(c)* The H_2 fraction remains roughly constant while the gas cloud cools (via the rotational and vibrational transitions of H_2) and condenses to higher density until it reaches a critical density and temperature at which gas cooling becomes inefficient. *(d)* Finally, runaway collapse begins once the dense clump's mass exceeds the local Jeans mass. Further fragmentation may then occur owing to gravitational instability, turbulence, or chemical processes.

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

where free electrons act as catalysts. We let the ionized fraction of hydrogen be $x_{\mathrm{H\ II}} = n_{\mathrm{H\ II}}/n_{\mathrm{H}}$, where $n_{\mathrm{H}} = n_{\mathrm{H\ I}} + n_{\mathrm{H\ II}}$ is the total abundance of hydrogen nuclei, and write the molecular fraction as $f_{\mathrm{H}_2} = n_{\mathrm{H}_2}/n_{\mathrm{H}}$. Then, considering only these two reactions and hydrogen recombination, we can write rate equations for a simplified reaction network as

$$\dot{x}_{\mathrm{H\ II}} = -\alpha_B n_{\mathrm{H}} x_{\mathrm{H\ II}}^2 \tag{5.3}$$

$$\dot{f}_{\mathrm{H}_2} = \tilde{k} n_{\mathrm{H}} (1 - x_{\mathrm{H\ II}} - 2 f_{\mathrm{H}_2}) x_{\mathrm{H\ II}}, \tag{5.4}$$

where the first equation follows recombinations (and hence the free-electron fraction), and the second includes the steps of molecular hydrogen formation, which occurs at a net rate $\tilde{k}$. This net rate coefficient includes both equation (5.1), whose rate we shall call k_2 (see further), and equation (5.2), whose rate we shall

Table 5.1 Important Reaction Rates for Hydrogen Species as Functions of Temperature T in Kelvin [with $T_\xi \equiv (T/10^\xi\,\mathrm{K})$]. In the text, we refer to these rate coefficients as k_i, where i refers to the appropriate line in this table. For a comprehensive list of additional relevant reactions, see Haiman, Z., Thoul, A. A., & Loeb, A., *Astrophys. J.* **464**, 523 (1996); Haiman, Z., Rees, M. J., & Loeb, A., *Astrophys. J.* **467**, 522 (1996); and Abel, T., Anninos, P., Zhang, Y., & Norman, M. L., *Astrophys. J.* **508**, 518 (1997).

Reaction		Rate coefficient ($\mathrm{cm^3\,s^{-1}}$)
(1)	$H^+ + e^- \rightarrow H + h\nu$	$8.40 \times 10^{-11} T^{-1/2} T_3^{-0.2} (1 + T_6^{0.7})^{-1}$
(2)	$H + e^- \rightarrow H^- + h\nu$	$1.65 \times 10^{-18} T_4^{0.76+0.15 \log T_4 - 0.033 \log^2 T_4}$
(3)	$H + H^- \rightarrow H_2 + e^-$	1.30×10^{-9}
(4)	$H^- + H^+ \rightarrow 2H$	$7.00 \times 10^{-7} T^{-1/2}$
(5)	$H_2 + e^- \rightarrow H + H^-$	$2.70 \times 10^{-8} T^{-3/2} \exp(-43{,}000/T)$
(6)	$H_2 + H^+ \rightarrow H_2^+ + H$	$2.40 \times 10^{-9} \exp(-21{,}200/T)$
(7)	$H_2 + e^- \rightarrow 2H + e^-$	$4.38 \times 10^{-10} \exp(-102{,}000/T) T^{0.35}$
(8)	$H^- + e^- \rightarrow H + 2e^-$	$4.00 \times 10^{-12} T \exp(-8{,}750/T)$
(9)	$H^- + H \rightarrow 2H + e^-$	$5.30 \times 10^{-20} T \exp(-8{,}750/T)$
(10)	$H + e^- \rightarrow H^+ + 2e^-$	$5.85 \times 10^{-11} T^{1/2}$ $\exp(-157{,}809.1/T)(1 + T_5^{1/2})^{-1}$

insert multiplication sign at line break / PE

call k_3. However, H^- is fragile and can be destroyed by CMB photons; we must therefore include a second channel in which the H^- does *not* lead to molecular hydrogen. This reaction occurs at a rate[2]

$$k_{11} \approx 0.114 T_\gamma^{2.13} \exp(-8{,}650\ \mathrm{K}/T_\gamma)\ \mathrm{s^{-1}}. \tag{5.5}$$

Thus, the net rate of H_2 formation is

$$\tilde{k} \approx k_2 \left[\frac{k_3}{k_3 + k_{11}/[(1 - x_{\mathrm{H\,II}}) n_{\mathrm{H}}]} \right], \tag{5.6}$$

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

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

First, note that the ionized fraction is independent of $f_{\mathrm{H_2}}$, since the electrons act only as catalysts. Thus, because $\dot{x}_{\mathrm{H\,II}} \propto x_{\mathrm{H\,II}}^2$, recombination will be very

i / PE

[i] Table 5.2 in §5.3.2 shows the same for deuterium-mediated reactions. Each line shows the rate coefficient for one particular reaction; in the text, we refer to these as k_i, where i refers to the appropriate line in Table 5.1. These should be included in detailed calculations but have only minor effects on the star formation picture described in this section.

slow, and the reservoir of electrons will remain substantial for long periods of time (much longer than the recombination timescale). The solution with constant T and $n_{\rm H}$ (i.e., the inefficient cooling limit) is

$$x_{\rm H\ II}(t) = \frac{x^i_{\rm H\ II}}{1 + t/t^i_{\rm rec}}, \tag{5.7}$$

where $x^i_{\rm H\ II}$ is the initial ionized fraction (taken from cosmological calculations, as in Figure 2.5), ~~and~~

$$t^i_{\rm rec} = (x^i_{\rm H\ II}\alpha_B n_{\rm H})^{-1} \approx 2.2 \times 10^8 \left(\frac{1+z}{20}\right)^{-3} \left(\frac{\Delta}{200}\right)^{-1} \left(\frac{x^i_{\rm H\ II}}{2\times10^{-4}}\right)^{-1} \text{yr} \tag{5.8}$$

is the recombination time at the initial ionized fraction, and $\Delta \equiv \rho/\bar{\rho}$. In the second part, we assumed the gas has an overdensity of ~200 (typical of virialized objects), we used the residual ionized fraction following recombination (Fig. 2.5), and we adopted a temperature $T \approx 10^3$ K.

We can now substitute this expression into equation (5.4). The factor $(1 - x_{\rm H\ II} - 2f_{\rm H_2})$ remains near unity for the initial conditions and timescales of interest. If we further approximate $\tilde{k}$ as roughly constant in this regime, the equation becomes integrable and yields

$$f_{\rm H_2} \approx f^i_{\rm H_2} + \frac{\tilde{k}}{\alpha_B} \ln(1 + t/t^i_{\rm rec}), \tag{5.9}$$

where $f^i_{\rm H_2}$ is the initial molecular fraction when the cloud forms (typically the IGM value after recombination, $\sim 6 \times 10^{-7}$, provided there is not yet a radiation background from luminous sources).[3] The molecular fraction therefore increases linearly with time when $t/t^i_{\rm rec} \ll 1$, but it slows to logarithmic growth past that point: the transition occurs when the electrons are incorporated into hydrogen atoms, which removes the population of catalysts and hence dramatically slows H_2 formation. It occurs at a critical molecular fraction

$$f_{\rm H_2,s} \equiv \frac{\tilde{k}}{\alpha_B} \approx 3.5 \times 10^{-4} (T/10^3\ {\rm K})^{1.52}, \tag{5.10}$$

where the s indicates saturation (though in actuality $f_{\rm H_2}$ does continue to increase slowly). In practice, the nominal recombination time at $\Delta \sim 200$ inside these objects is rather close to the Hubble time, so the electrons are used up quickly in the denser centers of the halos (with $\Delta \gg 200$), where molecule formation is also fastest. Thus most virialized objects reach this "saturation" limit after reasonably short timescales.

The upper two panels of Figure 5.3 illustrate this process in a numerical simulation of the formation of the first stars, a technique that has proved instrumental in understanding the process.[4] Panel a shows the free-electron fraction in a collapsing gas cloud as a function of density, but one can crudely also

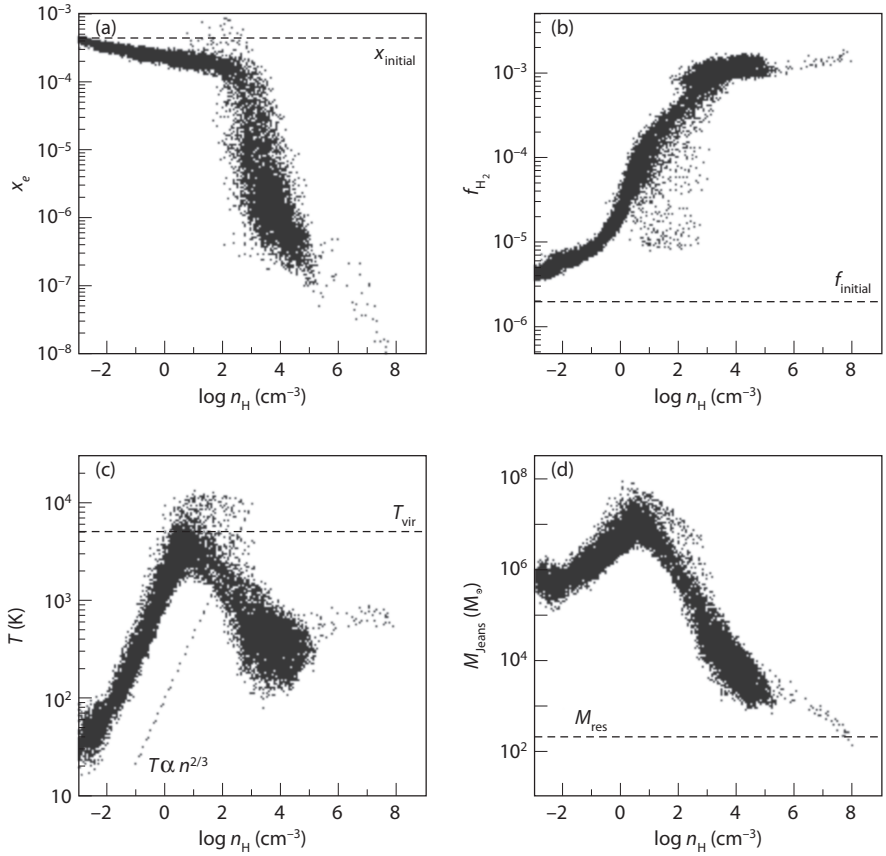


Figure 5.3 Gas properties during dense cloud collapse in a numerical simulation of first star formation. *(a)* Free-electron abundance; note the rapid decline at $n_H \sim 10^3$ cm^{-3}, owing to efficient recombinations. *(b)* Molecular fraction f_{H_2}. The fraction increases rapidly during cloud collapse until the saturation value (equation 5.10) is reached, when recombinations remove the free-electron catalysts. *(c)* Gas temperature as a function of number density. Note the strong clump at $T \sim 500$ K and $n_H \sim n_{cr}$, when radiative cooling becomes inefficient, so the evolution stalls. *(d)* The Jeans mass for this gas; note that $M_J \approx 10^3\ M_\odot$ for gas in the aforementioned stalling stage. Bromm, V., et al., *Astrophys. J.* **564**, 23 (2002). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~ by / AAS. /

consider it a function of time, since the density of a typical gas parcel will increase as it falls farther into the dark matter potential well. The electron fraction remains near the initial value (shown by the horizontal dashed line) for a period before falling rapidly at $n_H > 10^3$ cm^{-3}, where recombinations become efficient. Panel b shows the molecular fraction, which increases steadily at low densities (and therefore early times in the collapse process) before reaching a limiting value near $f_{H_2,s}$ in the densest part of the clump.

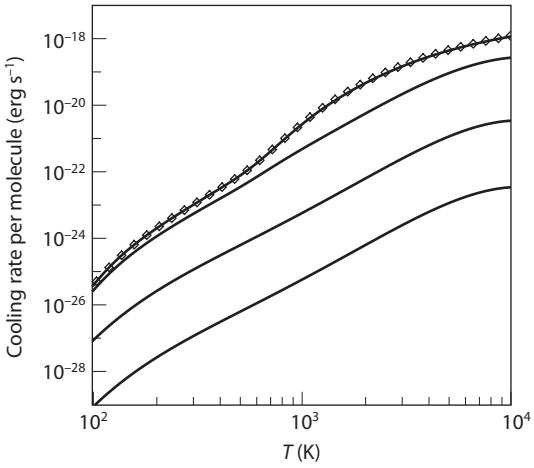


Figure 5.4 Cooling rate from H_2 per molecule. The solid lines show $n = 10^{-1}$, 10^{1}, 10^{3}, and 10^{5} cm^{-3}, from bottom to top. The diamonds show the cooling rate in LTE; note how the cooling function approaches this limit when $n > n_{cr}$ owing to the transition to LTE. Bromm V., et al., *Astrophys. J.* **564**, 23 (2002). Reproduced with permission of the American Astronomical Society. by / AAS. /

5.1.2 Cooling and Collapse of Primordial Gas

The tools in §5.1.1 allow us to follow the chemistry of the accretion of gas onto a virialized halo. The next question is, How much H_2 is required to allow the gas to cool and form stars. Cooling proceeds when an H_2 molecule is rotationally or vibrationally excited through a collision with another particle (see §6.1.1 for a detailed discussion of the energy levels of the H_2 molecule). If the subsequent de-excitation is radiative (and the cloud is optically thin), the cloud loses energy and cools; if it is de-excited through another collision, the cloud retains the energy, so no cooling occurs. In low-density gas, collisions are sufficiently rare that the first channel dominates, and the cooling rate is proportional to n_H^2 because all the molecules occupy low excitation states. Once collisions become important, the level populations shift to local thermodynamic equilibrium (LTE), and the cooling rate becomes proportional to n_H because the emergent intensity approaches the blackbody value. The transition occurs at the *critical density*, which is only a function of temperature; it corresponds to $n_{cr} \approx 10^4$ cm^{-3} for the temperatures of interest to primordial star formation. Figure 5.4 shows how the cooling rates depend on density and temperature: note how the higher density rates approach the LTE value near n_{cr}. The initial stages of cloud formation therefore lie in the low-density regime where cooling is efficient.

A halo can collapse from the overdensities characteristic of virialization to those characteristic of stars only if cooling can occur much faster than the timescale over which the halo grows (and therefore accumulates more thermal energy). The latter is comparable to the Hubble time. The cooling time depends on the reaction networks discussed in the previous section. But the

characteristic temperature to which H_2 radiation can drive gas is hundreds of Kelvins because the two lowest rotational energy levels in H_2 have an energy spacing of $E/k_B \sim 512$ K. A reasonable approximation to the cooling time in a virialized halo is[5]

$$t_{\rm cool} \approx 5 \times 10^4 f_{\rm H_2} \left(\frac{1+z}{20}\right)^3 \left(\frac{\Delta}{200}\right) \left(1 + \frac{10T_3^{7/2}}{60+T_3^4}\right)^{-1} \exp\left(\frac{512\ {\rm K}}{T}\right)\ {\rm yr}, \tag{5.11}$$

where $T_3 = T/(10^3\ {\rm K})$, and the temperature factors result from quantum-mechanical calculations of the H_2 collisional excitation rates.

The relevant comparison for determining whether a gas cloud will collapse rapidly to form stars is the dynamical time of the system, $t_{\rm dyn} \approx 1/\sqrt{G\rho}$ (with $\rho \sim m_p n_{\rm H}$), which describes how rapidly gravity can adjust the configuration of the system. If $t_{\rm cool} > t_{\rm dyn}$, the cloud can adjust to the (slow) cooling quasi-statically. It will contract slowly, maintaining a constant Jeans mass, so that $T \propto \rho^{1/3}$. If, however, $t_{\rm cool} < t_{\rm dyn}$, the gas cloud will lose all its thermal energy much faster than gravity can adjust the configuration. As the pressure support vanishes, the cloud will collapse to much higher densities in roughly the free-fall time.[ii]

In the present case, the relevant dynamical time is the Hubble time, $t_{\rm H}$, because the cooling begins as soon as the cloud reaches high densities (or over a virialization time). Even after the halo forms, it will continue to accept gas (and thermal energy) and grow over roughly the same timescale. From equation (5.11), the critical molecular fraction necessary for rapid cooling to occur is[6]

$$f_{\rm H_2,c} \approx 1.6 \times 10^{-4} \left(\frac{1+z}{20}\right)^{-3/2} \left(\frac{\Delta}{200}\right) \left(1 + \frac{10T_3^{7/2}}{60+T_3^4}\right)^{-1} \exp\left(\frac{512\ {\rm K}}{T}\right). \tag{5.12}$$

If a halo is able to form enough H_2 so that $f_{\rm H_2} > f_{\rm H_2,c}$, it will cool rapidly and form dense, highly molecular clouds. If not, it will remain a moderately dense, virialized clump until it can surpass that threshold. We term such clumps *minihalos*.

Figure 5.5 shows that detailed numerical simulations of the early stages of structure formation confirm this picture. Each circle represents a single virialized object in the simulation; the filled circles contain dense, cooling clouds, while the open ones do not. The dashed line shows the saturation limit for the molecular fraction, $f_{\rm H_2,s}$: clearly, the simulated halos lie remarkably close to this estimate, with the scatter likely due to variations in the accretion history of halos. The solid line shows the critical cooling threshold required at each virial temperature, $f_{\rm H_2,c}$. The intersection of these two curves determines the

[ii] This argument is much broader than this particular application: it provides a useful minimal criterion for galaxy formation in a wide range of contexts.

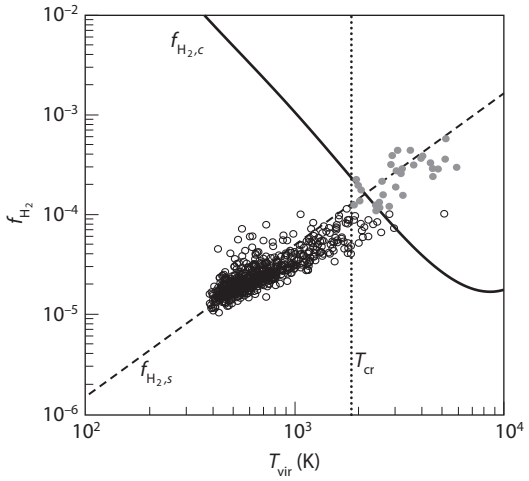


Figure 5.5 Molecular hydrogen fraction as a function of virial temperature for virialized halos inside a cosmological simulation at $z = 17$. The circles show results for individual halos; the filled circles contain dense (presumably star-forming) clouds, while the open circles do not. The dashed line shows the saturation limit $f_{H_2,s}$ of equation (5.10), while the solid line shows the critical molecular fraction $f_{H_2,c}$ for cooling to be rapid (see equation 5.12). The vertical dotted line shows the critical virial temperature for hosting star-forming clouds. Yoshida, N., et al., *Astrophys. J.* **592**, 645 (2003). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

critical threshold for star formation in molecular clouds, and the analytic argument provides a remarkably accurate criterion for determining which halos can host dense, star-forming clouds. As with Figure 5.5, the simplest way to think of this criterion is as a minimum mass threshold for cold cloud formation, because $f_{H_2,s}$ is an increasing function of temperature (and hence halo mass), while $f_{H_2,c}$ is a decreasing function of mass. The transition from minihalo to star-forming halo is therefore fairly sharp.

However, these simulations find that a fraction of halos lying above the nominal threshold still do not host star-forming clouds (the open circles in the upper right of Figure 5.5), while some lying below the nominal curve do have such clouds. These differences can be understood in terms of the accretion histories of the halos: recall that the cooling must balance the thermal energy gained throughout (ongoing) halo growth. Those halos accreting gas very rapidly may not be able to form dense clouds even if they are massive enough to have $f_{H_2,s} > f_{H_2,c}$.

DES: 2 lines under head ok?

5.1.3 The Collapse of Dense Clouds

Cloud collapse via H_2 radiation continues until cooling becomes inefficient and thermal pressure significant. The minimum temperature achievable by H_2

cooling is $T \sim 200$ K, because the energy spacing of the first two rotational levels of that molecule is ∼512 K (the limit is somewhat smaller than that nominal value because of the high-velocity tail of the Maxwell-Boltzmann distribution). The characteristic density when cooling becomes inefficient is the critical density $n_{\rm cr} \approx 10^4$ cm^{-3} defined in the previous section, where collisions become frequent enough to maintain local thermodynamic equilibrium. At yet higher densities, the radiative intensity must follow the blackbody law, so the cooling rate is only linearly proportional to density (see Figure 5.4).

With the decrease in the cooling rate, the gas cloud stalls or "loiters" at or near $n_{\rm cr}$. This stage is illustrated in panel c of Figure 5.3, which shows a phase diagram of the gas in a numerical simulation of cooling in high-z dark matter halos. In the early stages (i.e., gas at low density in this diagram), cooling is inefficient (with a rate proportional to $n_{\rm H}^2$), so the temperature roughly obeys the adiabatic relation $T \propto n_{\rm H}^{2/3}$ (shown by the dotted line here). Once the density increases enough for H_2 cooling to become efficient, the temperature falls to $T \sim 200$ K, where it stalls as LTE is reached near the critical density.

Further collapse requires enough mass to accumulate for gravity to overcome the roughly constant pressure of this growing clump—in other words, for the clump mass to exceed the local Jeans mass, $M_J \approx c_s t_{\rm coll}$ (see §3.2). For gas in this clump, that is

$$M_J \approx 700 \left(\frac{T}{200\ {\rm K}}\right)^{3/2} \left(\frac{n_{\rm H}}{10^4\ {\rm cm}^{-3}}\right)^{-1/2} M_\odot. \tag{5.13}$$

Once the clump grows beyond this point, gravity drives further, rapid collapse on the cloud's dynamical timescale $t_{\rm coll}$.

To this point in the collapse, "first star formation" poses a physics problem with well-specified initial conditions that can be solved on a computer. Starting with a simulation box in which primordial density fluctuations are realized (based on the initial power spectrum of density perturbations), one can reliably simulate the collapse by including the chemistry, gravitational dynamics, and thermodynamics of the gas. The top two panels in Figure 5.6 show these stages of collapse in a typical cosmological minihalo with $\sim 10^6 M_\odot$ in such a numerical simulation. Generically, the collapsing region makes a central massive clump with a typical mass of hundreds of solar masses, where the clump lingers because its H_2 cooling time is longer than its collapse time.

5.2 From Protostars to Stars

Although the journey that led to humanity's existence was long and complicated, one fact is clear: our origins can be traced to the production of the first heavy elements in the interiors of the first stars. Their formation is therefore a crucial milestone in the history of the Universe. The previous section has put us on the cusp of understanding these objects—but, unfortunately, the evolution from that point is much more difficult to understand and still has many uncertainties.

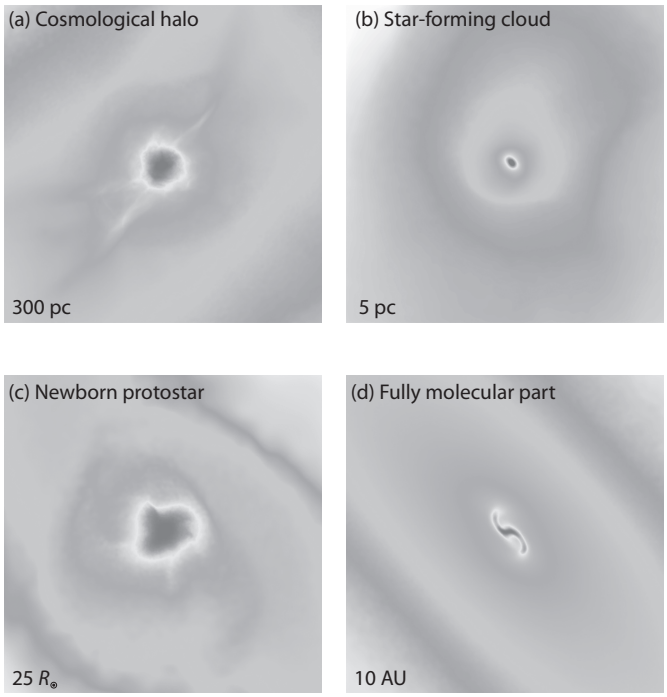


Figure 5.6 Projected gas distribution around a primordial protostar from a numerical simulation (see also *Color Plate 4* for a color version of this figure). Shown is the gas density of a single object on different spatial scales: *(a)* the large-scale gas distribution around the cosmological minihalo; *(b)* the self-gravitating, star-forming cloud; *(c)* the central part of the fully molecular core; and *(d)* the final protostar. Yoshida, N., Omukai, K., & Hernquist, L., *Science* **321**, 669 (2008). Copyright 2008 by the American Association for the Advancement of Science.

Numerical simulations show that the protostellar core, with $T \sim 200$ K, gradually contracts at roughly constant temperature (owing to H_2 cooling) until $n_H > 10^8$ cm^{-3}.[7] At that point, the density becomes large enough for three-body processes to form H_2 through the reactions

$$H + H + H \rightarrow H_2 + H, \tag{5.14}$$

$$H + H + H_2 \rightarrow H_2 + H_2. \tag{5.15}$$

Close up period-endnote / PE

The rate for the first of these reactions is $k_{3b} = 5.5 \times 10^{-29}$ cm^6 s^{-1}; the second is 1/8 as large.[8] The timescale for this reaction to proceed, $t_{3b} = (k_{3b} n_H^2)^{-1}$, equals the free-fall time at a critical density

$$n_{c,3} \approx \left(\frac{f_{H_2}^2 G m_p}{4 k_{3b}^2} \right)^{1/3}, \tag{5.16}$$

which is $\sim 10^8$ cm^{-3} for $f_{H_2} \sim f_{H_2,s}$. At higher densities, three-body H_2 formation proceeds rapidly, and the core collapses again. The molecular fraction then increases rapidly until it is near unity by the time $n_H \sim 10^{12}$ cm^{-3}, which one can estimate by setting $f_{H_2} \sim 0.5$ in equation (5.16). A simulated image of this stage is also shown in panel c in Figure 5.6.

At this point, the large molecular fraction rapidly increases the cooling rate, allowing dynamical collapse. Numerical simulations show that a hydrostatic core of mass $<10^{-2} M_\odot$ forms when the gas becomes optically thick to its own cooling radiation (panel d in Figure 5.6). This core forms the seed for a Population III star, but its subsequent evolution has proved much more difficult to predict robustly in numerical simulations. Not only is the dynamical time within the core very short, but radiative feedback from the protostar couples to the gas, making the cooling processes more complex. Thus, the final products of even the well-posed problem of Population III star formation still have a fair amount of uncertainty. Here, we content ourselves with identifying the key issues in these final stages of formation.

5.2.1 A Single Protostar: No Feedback

We begin by considering the simplest case, in which the clump is assumed to form a single protostar. Theorists have made a good deal of progress in understanding how such a protostar would grow using a combination of numeric and analytic tools.

Star formation typically proceeds from the inside out, through the accretion of gas onto a central hydrostatic core. Whereas the initial mass of the hydrostatic core is very similar for primordial and present-day star formation, the accretion process—ultimately responsible for setting the final stellar mass—is expected to be rather different. It is common to parameterize the accretion rate as

$$\dot{m}_\star = \phi_\star \frac{m_\star}{t_{\rm ff}}, \tag{5.17}$$

where $\phi_\star$ is a dimensionless parameter that depends on the properties of the medium, and $m_\star$ is the mass of the protostar. For a self-gravitating clump, the mass $m_\star \sim M_J \sim c_s^3/\sqrt{G^3\rho}$, the Jeans mass, so

$$\dot{m}_\star \sim c_s^3/G \propto T^{3/2}. \tag{5.18}$$

A simple comparison of the temperatures in present-day star forming regions, in which heavy elements cool the gas to a temperature as low as $T \sim 10$ K, with those in primordial clouds ($T \sim 200$–300 K), already indicates a difference in the accretion rate of more than two orders of magnitude. This suggests that the first stars were probably much more massive than their present-day analogs. The key factors to determine are the accretion rate itself and the duration over which it persists before radiative (or mechanical) feedback from the central protostar (or star) shuts it off.

To estimate the accretion rate quantitatively, we need to determine $\phi_\star$. The simplest interesting analog is spherically symmetric accretion in a uniform

medium onto a point mass, so-called Bondi accretion. A simple way to estimate how the accretion rate scales is to note that the protostar's gravity will overcome the pressure of the medium if the free-fall time $t_{\rm ff} \sim 1/\sqrt{G\rho}$ is smaller than the sound-crossing time $t_{\rm sc} \sim r/c_s$. This condition implies that infall will occur within a radius

$$R_{\rm acc} \sim \frac{Gm_\star}{c_s^2}. \tag{5.19}$$

The accretion rate will then be the surface area of a sphere at this radius, times the density of the medium, times the infall speed, which will be of the order of the sound speed. Thus,

$$\dot{m}_\star \sim \frac{G^2 m_\star^2 \rho}{c_s^3}. \tag{5.20}$$

We therefore have $\phi_\star \sim (m_\star/\rho r^3)(t_{\rm sc}/t_{\rm ff})^3 \sim 1$, as expected.

Population III star formation is of course considerably more complicated than this simplest limit, as collapse proceeds in a virialized clump and is regulated by H_2 cooling. Nevertheless, it is possible to estimate the rate of collapse by using the numerical simulations to calibrate more sophisticated models.[9] We take a *self-similar* solution, in which all relevant physical quantities are power laws, because there is no characteristic length scale in the problem. We assume that the density field follows $\rho \propto r^{-k_\rho}$ and that the pressure follows $p \propto r^{-k_p}$. It follows that the solution is a polytrope, with $p \propto \rho^{\gamma_p}$.

The simulations show that the accretion process occurs subsonically and nearly isentropically, with an adiabatic index $\gamma \approx 1.1$ set by the physics of H_2 cooling. In hydrostatic equilibrium, the configuration therefore assumes a polytropic solution with $P(r) \approx K\rho(r)^{1.1}$, so that $\gamma_p = 1.1$ as well. Moreover, hydrostatic equilibrium,

$$\frac{1}{\rho}\frac{dp}{dr} = -\frac{Gm_\star}{r^2}, \tag{5.21}$$

demands that $k_\rho = 2/(2-\gamma_p) \approx 20/9$ (i.e., the density structure is fairly close to an isothermal sphere), and $k_p = \gamma_p k_\rho$.

The constant K is set by the thermodynamics of the dense cloud during its "loitering" phase, which we can regard as the initial conditions of this stage of collapse. We take a fiducial value $K = 1.88 \times 10^{12} K_{\rm fid}$ in cgs units, where[iii]

$$K_{\rm fid} = \left(\frac{T}{300\ {\rm K}}\right)\left(\frac{10^4\ {\rm cm}^{-3}}{n_{\rm H}}\right)^{0.1}. \tag{5.22}$$

This initial entropy, together with the initial density profile, ultimately determines the accretion rate onto the protostar. The hydrostatic equilibrium condi-

[iii] In detail, the temperature here does not necessarily correspond to the gas temperature in the cloud if turbulence provides additional pressure support. We will consider the importance of turbulent motions later.

tion also requires that

$$\rho = \left[\frac{(3-k_\rho)k_p^3 K^3}{4\pi G^3 m_\star^2}\right]^{1/(4-3k_\rho)}, \tag{5.23}$$

Substituting into equation (5.17), we have

$$\dot{m}_\star = \frac{8\phi_\star}{\sqrt{3}}\left[\frac{(3-k_\rho)k_p^3 K^3}{2(2\pi)^{5-3\gamma_p} G^{3\gamma_p-1}}\right]^{1/2(4-3\gamma_p)} M^\xi, \tag{5.24}$$

$$\approx 0.026 K_{\rm fid}^{15/7}\left(\frac{m_\star}{M_\odot}\right)^{-3/7} M_\odot\ {\rm yr}^{-1}, \tag{5.25}$$

where $\xi = 3(1-\gamma_p)/(4-3\gamma_p)$. In the second line we have used $\gamma_p = 1.1$ and evaluated $\phi_\star$ using the closest known self-similar solution to the early stages of accretion in simulations.[10] There is, in fact, a fair amount of uncertainty in this relation because the exponent ξ (and hence the $m_\star$ dependence in equation 5.25) is very sensitive to γ_p, with the latter ranging from -0.37 to -0.49 for $\gamma_p = 1.09$–1.11; we have used $\gamma_p = 1.1$ here for concreteness. Nevertheless, the solution clearly shows an important fact—and a key difference from low-mass star formation—that the accretion rate actually tapers off with time. The time required to build up a given stellar mass is

$$t = \frac{m_\star}{\dot{m}_\star} \approx 27 K_{\rm fid}^{-15/7}\left(\frac{m_\star}{M_\odot}\right)^{10/7}\ {\rm yr}, \tag{5.26}$$

which matches detailed numerical simulations to within a factor of 2 or so in the early stages of protostar formation. Given that very massive Population III stars live for only a few million years, this provides a *maximal* upper limit to the mass of the final star of $\sim 10^3\ M_\odot$, the accumulated mass over that time period, which depends on both the main sequence lifetime and the initial entropy of the gas.

In detail, provided the core has some initial rotation, the gas falls onto an accretion disk rather than onto the star itself, and the resulting geometry may drive winds or other outflows, so the accretion rate estimated here is accurate only to a factor of the order of unity.

5.2.2 A Single Protostar: Radiative Feedback

The preceding maximal mass estimate assumes that the protostellar (and stellar) radiation field does not affect the accretion. In the presence of this feedback, *can a Population III star ever reach this asymptotic mass limit?* The answer to this question is not yet known with any certainty, and it depends on how that feedback manifests itself.

Before the onset of hydrogen fusion, the protostar must radiate away the gravitational energy accumulated by accretion, $L_{\rm acc} \approx G m_\star \dot{m}_\star / R_\star$, where $R_\star$ is its radius. The outward radiation pressure on the gas can itself halt accretion if it balances the inward gravitational force. This is the *Eddington luminosity* L_E,

which represents the maximal luminosity of an accreting object. Assuming for simplicity a fully ionized medium, force balance requires

$$\frac{Gm_\star m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{5.27}$$

where $\sigma_T = 0.677 \times 10^{-24}$ cm^2 is the Thomson cross section for scattering a photon off an electron. Setting $L_{\rm acc} \approx L_E$ yields a critical accretion rate,

$$\dot{m}_{\star,E} \approx \frac{L_E R_\star}{Gm_\star} \sim 5 \times 10^{-3} \left(\frac{R_\star}{5R_\odot}\right) M_\odot \ {\rm yr}^{-1}, \tag{5.28}$$

where we have scaled $R_\star$ to a value typical of a very massive Population III star on the main sequence. Comparison of equations (5.28) and (5.25) suggests that radiative feedback can be crucial in halting accretion onto the protostar as it approaches the main sequence with a mass ∼50–100 $M_\odot$.

However, radiative feedback is likely to be unimportant at much earlier stages, because the protostellar radii are much larger at these times. For example, in the very early stages, when the opacity is dominated by H^- bound-free processes, the photosphere temperature is fixed at $T \sim 6{,}000$ K because the opacity $\kappa_{H^-} \propto T^{14.5}$. Assuming that the protostar radiates as a blackbody, we then have

$$\frac{Gm_\star \dot{m}_\star}{R_\star} = 4\pi R_\star^2 \sigma_{\rm SB} T^4, \tag{5.29}$$

where $\sigma_{\rm SB}$ is the Stefan-Boltzmann constant. This expression yields $R_\star \approx 50\ (m_\star/M_\odot)^{1/3}\ R_\odot$ for $\dot{m}_\star \sim 0.005\ M_\odot\ {\rm yr}^{-1}$. Thus, we naively expect that radiative feedback will kick in only relatively late in the star formation process.

There are four distinct aspects of feedback exerted by a star on its gaseous environment:[11]

- *Photodissociation of H_2:* As the protostar heats up it produces UV radiation that photodissociates H_2 (see §6.1 for a detailed discussion). Once molecular cooling turns off, the adiabatic index of the gas increases to $\gamma = 5/3$ (i.e., monatomic gas). This decreases the accretion rate (because the pressure increases more rapidly as the gas gets compressed), but numerical estimates and semianalytic models show that the decline is rather modest. (This is not surprising given that the simple Bondi accretion problem described previously also permits steady accretion when $\gamma = 5/3$.)
- *Lyman-α radiation pressure:* As we will discuss in detail in §11.1.1, the radiative transfer of Lyman-α photons is typically a very complex process when the optical depth is very large, as occurs near a collapsing protostar surrounded by large quantities of neutral gas. The Lyman-α photons provide a substantial outward pressure, because they are trapped by the optically thick gas (and, on average, scattering off infalling gas blueshifts the photon, reducing the infall velocity of the gas). Indeed, they do not even escape by scattering through the gas column—rather, they escape when their frequency wanders so far from line center that the gas becomes effectively transparent. Because of these frequency shifts, the geometry of the flow plays an important role—as soon as a low–column density channel opens

up in one direction, photons can easily escape along that channel. Provided accretion occurs through a disk, Lyman-α escape is most likely to occur along the polar direction, where the accretion rate is already quite small. Analytic estimates show that Lyman-α scattering can begin to slow the accretion when the core has $M_c \sim 20\ M_\odot$ but that the overall effect is small. We consider this process in more detail in §6.3.2, where we examine its effects on larger scales.

- *Ionization:* Once the protostar begins to produce ionizing photons, it will carve out an H II region in which the temperature is much larger than the surrounding neutral gas (typically $>2\times10^4$ K; see §9.9 for a detailed discussion). This dramatically increases the pressure of the gas, which can cause the H II region to expand and drive off gas that would otherwise accrete onto the protostar. The dynamics of the region depend on the expansion velocity of the ionization front. If the front moves faster than about twice the ambient sound speed (of the neutral gas), then it has essentially no dynamical effect on the gas. This is known as an "R type" (or rarefied) front. Near a Population III protostar, the H II region begins in this regime, because it is expanding through gas falling inward at the free-fall velocity $v_{\rm ff}$, which is highly supersonic.

  Eventually, the front reaches the radius where $v_{\rm ff} \sim 2c_s$, where the gas can respond to the ionization front, and a shock forms (this is a "D type," or dense front). Typically, the shock leads the ionization front, creating a dense shell of neutral gas into which the front propagates, with a bulk kinetic energy density comparable to the pressure inside the ionization front. A simple estimate for the point at which this shock halts accretion is thus when the thermal pressure gradient at the front exceeds the inward gravitational force. This is roughly the accretion radius $R_{\rm acc}$ defined in equation (5.19), but using $T \sim 20{,}000$ K for the ionized gas. Estimates of the ionizing luminosity of these protostars indicate that this limit is reached when $m_\star \sim 100\ M_\odot$.

  As before, the disk geometry of the accretion flow plays an important role in how this feedback mechanism occurs. The front will propagate fastest through the lowest column density of gas, which is along the polar axis, so accretion will first be suppressed there. In contrast, along the direction of the disk, the extreme column density of the disk "shadows" the flow, allowing accretion to continue. Provided most of the accretion occurs through such a disk, the H II region will therefore not entirely halt the protostar's growth.

  Photoionization heating can have substantial effects on the gas even far outside the protostar's immediate environment. We examine this in more detail in §6.3.1.
- *Photoevaporation of the accretion disk:* However, the same ionizing photons heat the disk itself, evaporating gas from it and eventually shutting off accretion entirely. The rate at which this occurs depends on the geometry of the disk and the spectrum of the protostar, but some calculations show that the disk evaporates when $m_\star \sim 150\ M_\odot$. As we will see this is very near the

mass threshold for direct black hole formation when such stars die, so the details of the process may be very important.

Because these radiative feedback processes affected accretion only late in the evolution of the first stars, they must generally be studied with simplified analytic models rather than incorporated directly into ab initio simulations of Population III star formation. We therefore have only approximate estimates of their importance, and observations of these stars may be necessary to settle the physical uncertainties.

5.2.3 Multiple Protostars: Fragmentation

The models described in §5.2.2 make one key assumption: that the collapsing material accretes onto a single object, the central protostar. However, in the presence of angular momentum the accretion flow generically organizes itself into a disk. *Can this disk then fragment into multiple high-density clumps, or multiple protostars?* There are several possible mechanisms for fragmentation—gravitational instabilities, turbulence, and thermodynamic instabilities. All have now been implicated in numerical simulations showing fragmentation, but it is far from clear whether these are generic processes, or how severe the fragmentation is.

The classic way to gauge the importance of gravitational instability is the *Toomre criterion*. We sketch its significance here. Consider a small patch inside a rotating gaseous disk. Let the patch have a radius r and mass $M = \pi\Sigma r^2$ (where $\Sigma = \rho/\Delta z$ is the surface density, and Δz is the disk thickness). If we compress the patch by a factor δ, so $r \rightarrow r(1-\delta)$, the pressure increases by an amount

$$\Delta p \sim c_s^2 \delta\rho_0 \sim \delta c_s^2 \Sigma (\Delta z)^{-1}. \tag{5.30}$$

Thus, the excess pressure force per unit mass is

$$\frac{\nabla(\Delta p)}{\Sigma(\Delta z)^{-1}} \sim \frac{c_s^2\delta}{r}, \tag{5.31}$$

where we have assumed that r is the characteristic scale over which the system varies. Meanwhile, the increase in the gravitational force per unit mass is $-GM\delta/r^2 \sim G\Sigma\delta$. Thus, the outward pressure counteracts gravity if

$$r < \frac{c_s^2}{G\Sigma} \equiv R_{\rm pr}. \tag{5.32}$$

This is just the classical Jeans analysis (§3.2) applied to a two-dimensional system: small-wavelength modes are stabilized by pressure, while large-wavelength modes are unstable to gravitational collapse.

However, in a rotating disk the angular momentum can stabilize these long-wavelength modes. Assuming that our perturbation involved no external force (and hence torque), the internal spin angular momentum (generated by differential rotation across the patch) must be conserved. If Ω is the rotation speed, this is $J_s \sim \Omega r^2$.

As we compress the patch, conservation of angular momentum increases the rotation speed and thus creates a centrifugal barrier to further compression. To gauge how effective this barrier is, we write the centripetal force per unit mass in terms of the conserved quantity J_s:

$$\frac{v^2}{r} \sim \frac{\Omega^2 r^2}{r} \sim \frac{J_s^2}{r^3}. \tag{5.33}$$

Thus, the excess force as we compress the patch is $d(J_s^2/r^3)/dr \times \delta r$, which overcomes gravity and prevents further collapse if

$$r > \frac{G\Sigma_0}{\Omega^2} \equiv R_{\rm cen}. \tag{5.34}$$

We can have an instability only if $R_{\rm cen} > R_{\rm pr}$, in which case there exists a range of moderate-wavelength perturbations that cannot be stabilized by either pressure or rotation. A more exact derivation shows that instability sets in if the *Toomre criterion*

$$Q \equiv \frac{c_s \kappa_e}{\pi G \Sigma} < 1. \tag{5.35}$$

Here κ_e is the epicycle frequency, or the rotation frequency for small perturbations around the equilibrium disk. For a Keplerian disk, $\kappa_e = \Omega = \sqrt{GM(r)/r}$, where $M(r)$ is the mass enclosed within a radius r. If the disk is unstable, fragmentation will generically occur as positive density perturbations grow rapidly.

change to `epicyclic' /AA

Delete comma / AA

Figure 5.7 shows this kind of gravitational fragmentation in a numerical simulation of the accretion disk around a Population III star.[12] The disk very quickly exhibits spiral structure—common in self-gravitating disks—and develops nonaxisymmetric features and becomes locally unstable just ∼100 years after the formation of the first protostellar core. The disk soon forms a second core separated by ∼20 AU from the first. Figure 5.8 shows why: the top two panels show that the surface density and temperature of the disk remain roughly constant over time, except near its outskirts. This means the rate at which the disk can transport angular momentum (and hence material) inward stalls, and the outer disk builds up more and more mass, quickly becoming gravitationally unstable ($Q \sim 1$ at $r \sim 20$ AU).

To continue fragmentation, the clump must still be able to rid itself of the thermal energy generated during collapse. At the characteristic densities of these disks ($n_{\rm H} \sim 10^{12}$–10^{14} cm^{-3}) a new cooling process dominates: *collision-induced emission* (CIE). This occurs when H_2 interacts with another species (H, He, or H_2) in a collision. The interacting pair briefly forms a "supermolecule" with a nonzero electric dipole, from which photons can be emitted or absorbed efficiently. Because the collision times are very short, the uncertainty principle demands that the resulting radiation be emitted nearly in a continuum. This CIE radiation allows the gas to cool during the early stages of fragmentation, because the cooling time is substantially shorter than the dynamical time.

The continuum opacity of these same molecules prevents CIE cooling at $n_{\rm H} > 10^{16}$ cm^{-3}. At this point, the gas does begin to heat up. However, at temperatures much above the $T \sim 10^3$ K characteristic of the disk (see the upper

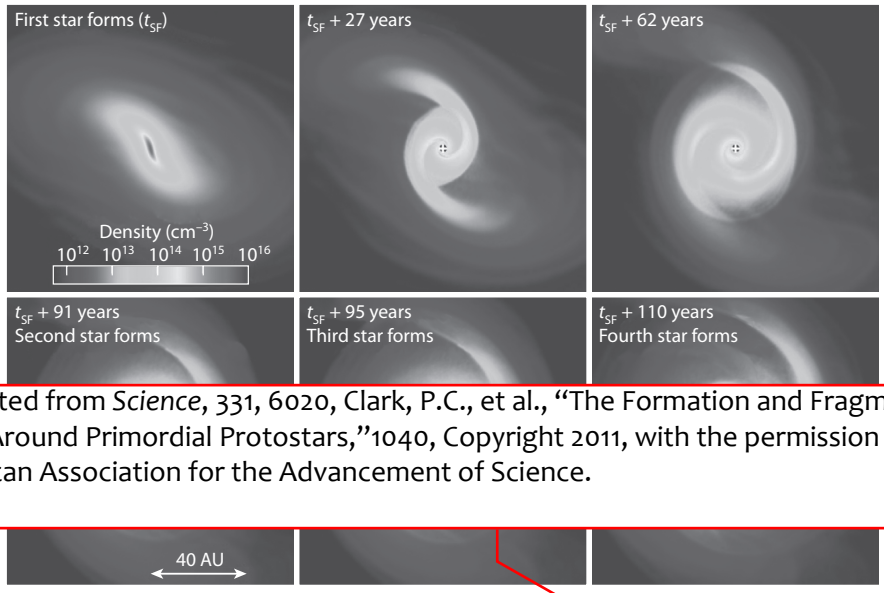


Figure 5.7 Density evolution in a 120 AU region around the first protostar in a numerical simulation of Population III star formation, showing the buildup of the protostellar disk and its eventual fragmentation at the times labeled in the diagram (see *Color Plate 5* for a color version of this figure). ~~Clark, P. C., et al., *Science* 331, 1040 (2011). Copyright 2011 by the American Association for the Advancement of Science.~~

right panel in Fig. 5.8), H_2 begins to dissociate. Each such dissociation removes 4.48 eV from the gas, which keeps it near its original temperature because it begins so highly molecular (see the lower left panel in Figure 5.8).

Turbulence appears to be a third factor triggering instabilities and fragmentation. Such turbulence can be generated by "cold" accretion onto the host minihalo, where gas is funneled into the halo along filamentary channels and is not initially shock heated to the virial temperature of the halo. Instead, it collides with the central gas clump supersonically, triggering (typically subsonic) turbulent motions. Turbulence is known to be important in "normal" star formation at low redshifts, and leads to fragmentation of giant molecular clouds into protostellar cores with a wide range of initial masses. Some numerical simulations indicate that similar processes could cause fragmentation in the Population III regime.[13]

5.2.4 The Initial Mass Function

Currently, we have no direct observational constraints on how the first stars formed at the end of the cosmic dark ages, in contrast to the wealth of observational data we have on star formation in the local Universe.[14] Population I and II stars form out of cold, dense molecular gas that is structured in a

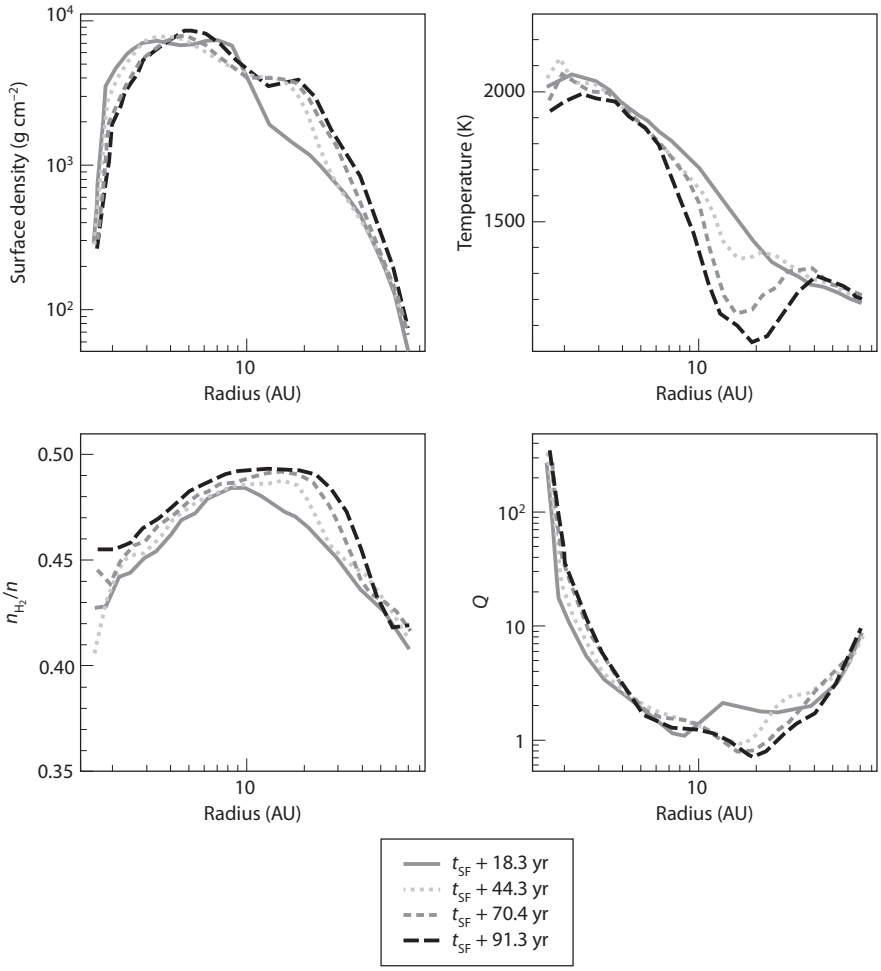


Figure 5.8 Radial profile of disk physical properties from the same simulation shown in Figure 5.7, centered on the first protostellar core to form. Clockwise from upper left, the panels show the surface density, temperature, Toomre Q-parameter, and molecular fraction. Note that the disk parameters do not evolve strongly with time (shown with the different curves in each panel). The second core to form in the simulation forms within the region near $r \sim 20$ AU where $Q < 1$. ~~Clark, P. C., et al., *Science* **331**, 1040 (2011). Copyright 2011 by the American Association for the Advancement of Science.~~

complex, highly inhomogeneous way. The molecular clouds are supported against gravity by turbulent velocity fields and are pervaded by magnetic fields. Stars tend to form in clusters, ranging from a few hundred up to $\sim 10^6$ stars. It appears likely that the clustered nature of star formation leads to complicated dynamical interactions among the stars. The *initial mass function* (IMF) of

Population I stars is observed to have a broken power-law form, originally identified by Ed Salpeter,[15] with a number of stars $N_\star$ per logarithmic bin of stellar mass $m_\star$,

$$\frac{dN_\star}{d\log m_\star} \propto {m_\star}^{-\Gamma}. \tag{5.36}$$

Figure 5.9 shows some data in nearby star-forming regions, the only environment in which the IMF can reliably be measured, and the effective power-law index in these regions. The data are consistent with a broken power law,

$$\Gamma \simeq \begin{cases} 1.35 & \text{for } m_\star > 0.5\,M_\odot \\ 0.0 & \text{for } 0.008\,M_\odot < m_\star < 0.5\,M_\odot \end{cases}. \tag{5.37}$$

We take this as our fiducial model in the discussion, though we note that the form of the IMF at low masses is still unsettled. The lower cutoff in mass corresponds roughly to the minimum fragment mass, set when the rate at which gravitational energy is released during the collapse exceeds the rate at which the gas can cool.[16] Moreover, nuclear fusion reactions do not ignite in the cores of protostars below a mass of $\sim$0.08 $M_\odot$, so-called brown dwarfs. The most important feature of this IMF is that $\sim$0.5 $M_\odot$ characterizes the mass scale of Population I and II star formation, in the sense that most of the stellar mass goes into stars with masses close to this value.

The ultimate goal of studies of the formation of Population III stars is to determine the analogous mass function for primordial stars. Unfortunately, we are far from converging on any robust predictions. Until recently, models of single protostar formation seemed to suggest that accretion would continue until $m_\star \sim 100\ M_\odot$, with the details determined by the initial entropy of the gas ($K_{\rm fid}$ in equation 5.25) and by radiative feedback, with a plausible mass range from $\sim$20 to 300 $M_\odot$. These masses—obviously much larger than the characteristic mass of present-day stars—suggested that the first generation of stars to light up the Universe would truly have been exotic objects.

However, the more recent studies of fragmenting disks suggest that the characteristic masses may be much smaller. Gravitational instability leads to several cores, each competing for the accreting gas. Turbulence may lead to an even wider range of initial protostar sizes. These cores themselves can interact, much as the stars in nearby open clusters do. In particular, three-body interactions tend to speed up smaller cores and move them into the outskirts of the core, where there is less gas to accrete. Meanwhile, the larger cores tend to sink to the center of the cloud, accreting more rapidly. This picture of "competitive accretion" may be important for high-mass star formation in the nearby Universe; if so, it may suggest that Population III star formation may also follow a power-law IMF with a broad range of stellar masses.

Nevertheless, it seems likely that the characteristic mass of high-redshift stars *must* be significantly larger than the present-day value of $\sim$0.5 $M_\odot$. The present-day value can be understood relatively easily as the minimum mass for collapse in the $\sim$8 K molecular gas out of which these stars form (the minimum temperature is set by the cooling physics in molecular clouds). The Jeans mass

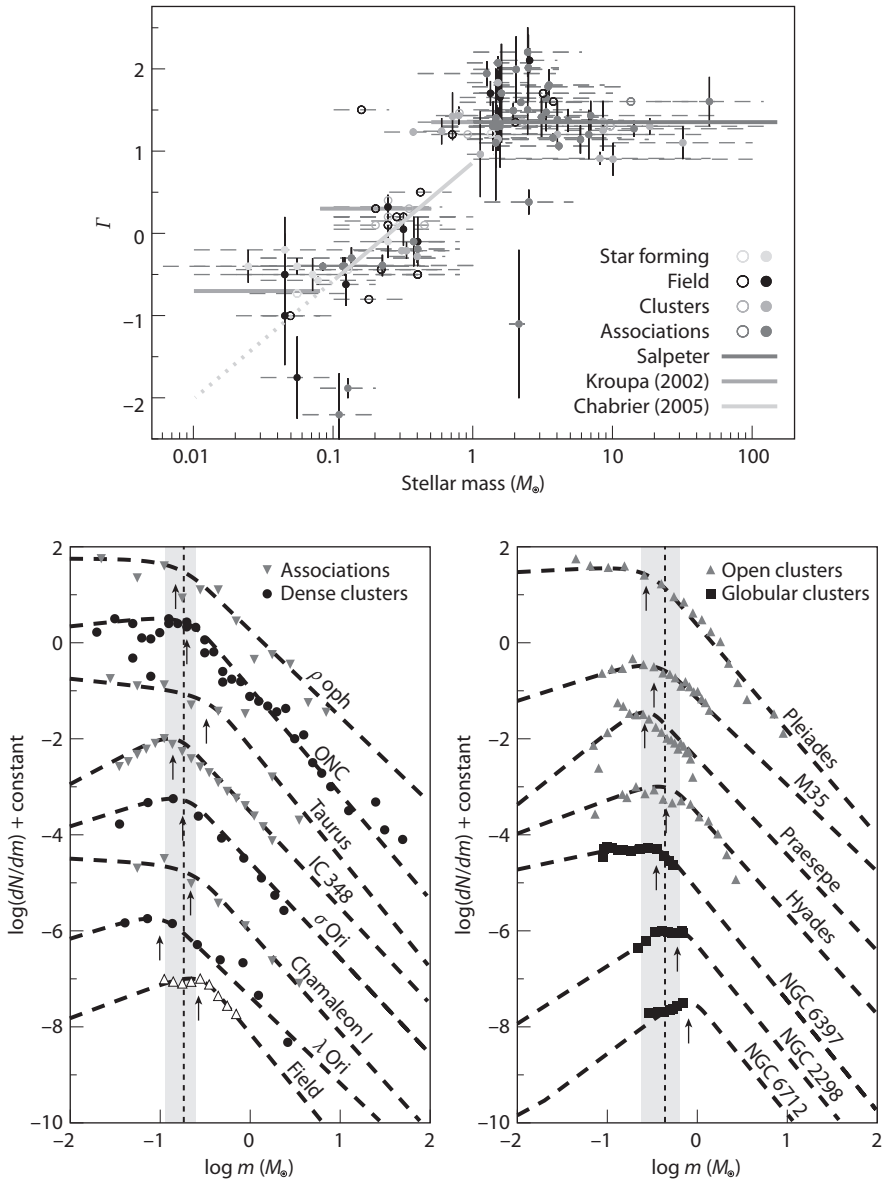


Figure 5.9 *Upper panel:* The derived power-law index, Γ, of the IMF in nearby star-forming regions, clusters, and associations of stars within the Milky Way galaxy, as a function of sampled stellar mass (points are placed in the center of the log $m_\star$ range used to derive each index; the dashed lines indicate the full range of masses sampled). The ~~colored~~ solid lines represent three analytic 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, the dotted line indicates the mean

AA

provides a reasonable estimate of this value, but a more exact choice is the *Bonnor-Ebert mass,*[17]

$$M_{\rm BE} = 1.18 \frac{(k_B T/\mu m_p)^2}{p_0^{1/2} G^{3/2}}, \tag{5.38}$$

which is the largest mass that an isothermal gas sphere with a temperature T can have in hydrostatic equilibrium with an external gas pressure p_0. A Bonnor-Ebert sphere has a finite central density and size, as it is confined by external pressure. Its maximum mass $M_{\rm BE}$ is 4.7 times smaller than the Jeans mass but otherwise has the same scaling with density and temperature.

The temperature floor is expected to evolve with redshift, because radiative cooling cannot bring the temperature below the CMB temperature, to which all the relevant lines couple. At $z = 30$, $T_{\rm CMB} = 82$ K, many times larger than the present-day value (which is actually well above the $z = 0$ CMB temperature). The quantitative change in the Bonner-Ebert mass is not trivial to estimate, because it depends on the temperature–density relation in the clouds: for example, if the density structure is fixed, $M_{\rm BE} \propto T^{3/2}$, but if cooling proceeds isobarically, with $nT =$ constant, then $M_{\rm BE} \propto T^2$. This suggests that the characteristic fragmentation mass would increase to at least ~16–50 $M_\odot$ at $z = 30$ (or even 10–20 $M_\odot$ at $z = 10$), well into the range of "high-mass" stars by present-day standards, though far smaller than the maximal estimates if fragmentation is inefficient.

We discuss the IMF in a more general context in §8.8 as well.

5.3 The Second Generation of Stars: "Population III.2"

The picture we have described so far assumes that the star formation process begins with the initial conditions characteristic of the high-redshift IGM: gas that is nearly neutral, with very little preexisting H_2. These were, of course, the proper initial conditions for the first star-forming halos. But this picture depends rather sensitively on those assumptions, and it is likely that later generations of stars—still forming out of primordial gas— began with very different conditions.

The key is the initial ionization state of the gas. There are three important ways in which that could have been much higher for these later stars. One possibility is that the first stars produced a copious amount of ionizing radiation, generating H II regions within and around their host dark matter halo.

Figure 5.9 (Continued.) characteristic mass among the clusters in each panel, and the shaded region shows the standard deviation of the characteristic masses in that panel. The observations are consistent with a single underlying IMF. (See *Color Plate 6* for a color version of this figure.) Bastian, N., Covey, K. R., & Meyer, M. R., *Ann. Rev. Astr. & Astrophys.* 48 (2010). Copyright 2010 by Annual Reviews.

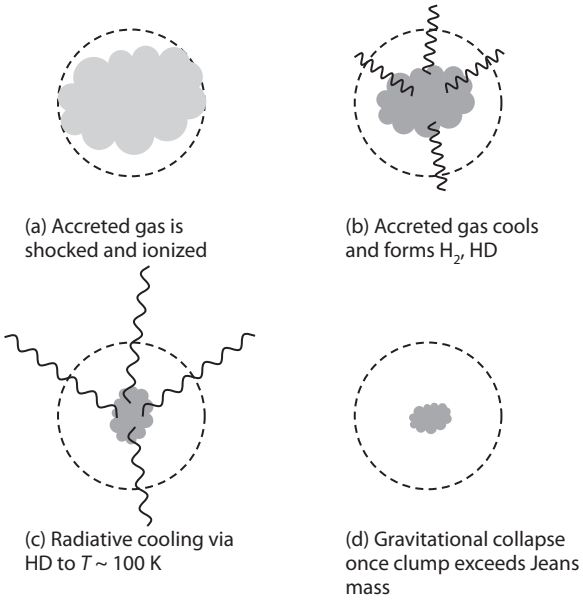


Figure 5.10 Basic stages in Population III.2 protostar formation. *(a)* A dark matter halo passes the threshold to ionize infalling gas. *(b)* The accreted gas cools rapidly owing to H I line cooling. Because the gas is initialy highly ionized, large fractions of H_2 and HD are formed in the process. *(c)* Instead of stalling, HD line cooling continues to function to $T \sim 100$ K, substantially below the "loitering" temperature for Population III.1 star formation. *(d)* Finally, runaway collapse begins once the mass of the dense clump exceeds the local Jeans mass. Further fragmentation may then occur owing to gravitational instability, turbulence, or chemical processes.

Any clumps that collapsed within the ionized region would have collapsed from fully ionized gas. Similarly, if these stars exploded in supernovae, their powerful blast waves would have ionized the nearby gas (and possibly even triggered collapse). Finally, as larger halos formed, star formation would have shifted to those more massive objects. Above a virial temperature of $\sim 10^4$ K, the virialization shock itself would ionize the halo gas, and again change the initial conditions for cloud chemistry and collapse.

These initial conditions resulted in a different formation mode for primordial stars, often referred to as *Population III.2*, with an initial mass function distinct from the classic Population III.1 mode described earlier. The basic stages in this process are illustrated in Figure 5.10.

COMP: Poor wordspacing. / PE

DES: 2 lines under head ok?

5.3.1 The Freeze-Out of Molecular Hydrogen

We showed in §5.1.1 that H_2 formation is catalyzed by the presence of free electrons. Thus, in gas that cools from a fully ionized state, molecule

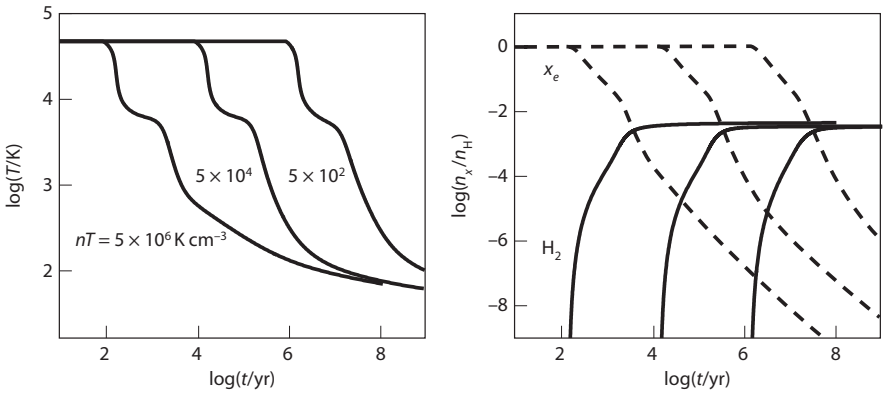


Figure 5.11 H_2 formation in initially ionized gas. The left panel shows the temperature evolution of gas at three different initial pressures, assuming isobaric cooling (the three models are offset in time for clarity of presentation). The right panel shows the molecular fractions (solid curves) and the free electron fractions (dashed curves) for the same three models. Note that f_{H_2} approaches a constant limit in all three cases. Oh, S. P., & Haiman, Z., *Astrophys. J.* **569**, 558 (2002). Reproduced with permission of the American Astronomical Society.

formation can proceed rapidly—even though at the initially high temperatures such molecules quickly dissociate.

Figure 5.11 shows numerical models of idealized isobaric cooling in primordial gas initially at $T \sim 10^4$ K (and hence ionized). As the gas cools, H_2 begins to form through the usual free-electron channel, until its abundance saturates at $f_{H_2} \sim 2 \times 10^{-3}$, regardless of the initial conditions. This "freeze-out" level indicates that the molecular fraction saturates at a nonequilibrium value.

In particular, f_{H_2} can no longer evolve once the timescales for H_2 formation (t_{form}) and dissociation (t_{diss}) become longer than the cooling and recombination timescales in the system, because the electron catalysts disappear at that point. As in §5.1.1, the formation time can be approximated by $t_{form} = f_{H_2}/\dot{f}_{H_2} \approx f_{H_2}/(x_{H\ II}\tilde{k}n_H)$. The dominant H_2 dissociation process is reaction 6 in Table 5.1, whose rate we denote by k_6. (The dissociation chain beginning with this charge-exchange reaction is more efficient than direct collisional dissociation.) Then, $t_{diss} = (k_6 x_{H\ II} n_H)^{-1}$. The rate $t_{diss}^{-1} \propto k_6$ decreases exponentially as the temperature drops, while the rates for cooling, recombination, and formation decrease only as power laws. This steep temperature dependence means that t_{diss} very suddenly becomes longer than t_{rec} and t_{cool} as the gas cools; the reaction rates demand that the resulting temperature is $T_{freeze} = 3{,}700$ K.[18] Up to this point, the H_2 abundance remains in equilibrium, and the ratio of the reaction rates yields the value

$$f_{H_2,freeze} \approx \frac{\tilde{k}(T_{freeze})}{k_6(T_{freeze})} \approx 2 \times 10^{-3}. \quad (5.39)$$

This argument shows that at lower temperatures, molecular hydrogen will no longer be destroyed; thus at lower temperatures we must have $f_{H_2} > f_{H_2,\mathrm{freeze}}$. Furthermore, at T_{freeze} equilibrium demands that the formation and dissociation timescales be comparable to each other and to t_{rec} and t_{cool}. To show that the molecular fraction does not increase above this freeze-out value, we need only to verify that t_{form} is longer than t_{rec} and t_{cool} at lower temperatures, which is straightforward. This further implies that H_2 formation will also cease so long as its rate of formation increases less slowly with temperature than its rates of cooling and recombination, which can readily be seen by comparing the reaction rates in Table 5.1. Thus, when $T < T_{\mathrm{freeze}}$, the molecular hydrogen abundance remains fixed at its (nonequilibrium) freeze-out value $f_{H_2,\mathrm{freeze}}$.

5.3.2 Deuterium and Cooling

The relatively high abundance of molecular gas already suggests that these preionized systems can also eventually cool and form stars. However, there is an additional wrinkle that becomes important in these systems: deuterium. Unlike H_2, which is a symmetric molecule, HD has a permanent dipole moment, which allows strong dipole rotational transitions, with $\Delta J = \pm 1$, of smaller energy than the $\Delta J = \pm 2$ quadrupole transitions of H_2 (the larger reduced mass of HD lowers this energy even further). The $J = 1 \rightarrow 0$ transition has an equivalent temperature of $\sim$130 K, about four times smaller than the lowest energy transition of H_2. Thus, in principle, HD cooling can lower the temperature and hence mass scale of star formation substantially (recall that $M_J \propto T^{3/2}$ at fixed density, equation 5.13).

The most efficient method for HD to form is via the reaction

$$H_2 + D^+ \rightarrow HD + H^+, \tag{5.40}$$

which, of course, requires the simultaneous presence of molecular hydrogen and ionized deuterium. In the standard Population III.1 picture, which occurs entirely at low temperatures, the latter is very rare, and very little HD forms. However, in the present case, where all the deuterium is initially ionized, the abundance of D^+ remains relatively large until very low temperatures. Thus, a substantial abundance of HD can build up, as illustrated in Figure 5.12. Because the Big Bang nucleosynthesis expectation is that the deuterium abundance is only $\sim 10^{-5}$ that of hydrogen, these calculations indicate that nearly all the deuterium can enter molecular form. Table 5.2 provides reaction rates for the most important deuterium reactions.

Moreover, HD has several advantages as a coolant over H_2. First, it has a higher critical density, $n_{\mathrm{crit,HD}} \sim 10^6$ cm^{-3}, so rapid cooling continues to higher densities. Second, its dipole transitions are much more rapid, with a spontaneous decay rate $A_{10} \approx 5 \times 10^{-8}$ s^{-1}. This allows rapid cooling even at low abundances: at the levels shown in Figure 5.12, the gas can easily cool to the CMB temperature over a relatively short time. To see this, let us assume for simplicity that the gas, at temperature T, is in LTE, so that the population levels

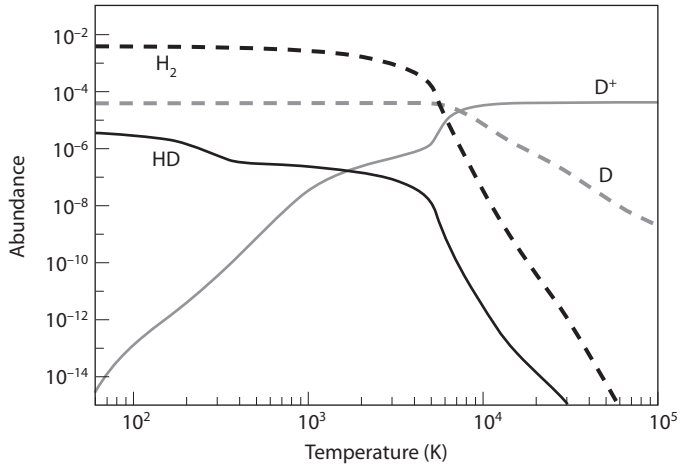


Figure 5.12 Molecular abundances in primordial gas cooling from high temperatures, relative to the total number density of H atoms. The calculation here simulated cooling in a 100 km s^{-1} shock at $z = 20$, characteristic of a supernova. Note the large abundance of HD at low temperatures. Johnson, J. L., & Bromm, V., *Mon. Not. R. Astron. Soc.* **366**, 247 (2006). Copyright 2006 by the Royal Astronomical Society.

Table 5.2 Reaction Rates for Deuterium Species as Functions of Kelvin Temperature T [with $T_\xi \equiv (T/10^\xi\ \mathrm{K})$]. Adopted from Haiman, Z., Thoul, A. A., & Loeb, A., *Astrophys. J.* **464**, 523 (1996); and Galli, D., & Palla, F., *Astron. & Astrophys.* **335**, 403 (1998).

Reaction		Rate coefficient ($cm^3\ s^{-1}$)
(1)	$D^+ + e^- \rightarrow D + h\nu$	$8.40 \times 10^{-11} T^{-1/2} T_3^{-0.2} (1 + T_6^{0.7})^{-1}$
(2)	$D + H^+ \rightarrow D^+ + H$	$3.70 \times 10^{-10} T^{0.28} \exp(-43/T)$
(3)	$D^+ + H \rightarrow D + H^+$	$3.70 \times 10^{-10} T^{0.28}$
(4)	$D^+ + H_2 \rightarrow H^+ + HD$	2.10×10^{-9}
(5)	$HD + H^+ \rightarrow H_2 + D^+$	$1.00 \times 10^{-9} \exp(-464/T)$

in the ground (n_0) and first excited state (n_1) are

$$\frac{n_1}{n_0} = 3e^{-T_D/T}, \tag{5.41}$$

where the ratio of statistical weights is 3, and $T_D = h\nu_{10}/k_B \approx 130$ K is the equivalent temperature for photons emitted in transitions from the first excited state to the ground state. (We take a two-level system for simplicity, assuming that the gas has already cooled to $T \sim T_D$, so that higher levels are rare.)

AA

The radiative cooling rate of gas at constant density is

$$h\nu_{10}(n_0 B_{01} I_{\nu_{10}} - n_1 A_{10} - n_1 B_{10} I_{\nu_{10}}) = \frac{3}{2} n k_B \frac{dT}{dt}. \quad (5.42)$$

Here, B_{01} and B_{10} are the Einstein coefficients for stimulated emission and absorption, respectively; and n is the *total* number density of particles: this value is related to the density of HD molecules $n_{\rm HD} = n_0 + n_1$ by $X_{\rm HD} = n_{\rm HD}/n$. Finally, $I_{\nu_{10}}$ is the CMB intensity at the frequency of the HD $J = 1 \rightarrow 0$ transition,

$$I_{\nu_{10}} \approx \frac{2h\nu_{10}^3}{c^2} e^{-T_D/T_{\rm CMB}} = \frac{A_{10}}{B_{10}} e^{-T_D/T_{\rm CMB}}, \quad (5.43)$$

where we have used the fact that $T_D \gg T_{\rm CMB}$. In that case the stimulated emission term can also be neglected, so equation (5.42) may be written

$$\frac{dT}{dt} \approx 2T_D A_{10} X_{\rm HD} (e^{-T_D/T_{\rm CMB}} - e^{-T_D/T}). \quad (5.44)$$

If we assume that $X_{\rm HD}$ remains constant, we can integrate this equation to find that the time to cool from $T \sim T_D$ to $T = T_{\rm CMB}$ is[19]

$$t_{\rm HD,cool} \sim 1/(X_{\rm HD} A_{10}), \quad (5.45)$$

at $z \sim$ 10–30. Equating this result to the Hubble time, we can determine the critical HD abundance for cooling as

$$X_{\rm HD,crit} \sim 4 \times 10^{-9} \left(\frac{1+z}{30}\right)^{3/2}. \quad (5.46)$$

Figure 5.12 (and similar calculations for other scenarios) show that when cooling from high temperatures, the gas forms far more HD than this critical value, which implies very efficient HD cooling. However, the abundance of HD in the "normal" Population III.1 scenario is well below this critical value—because D^+ is so rare in cold gas—so it is not an important coolant for that star formation channel.

5.3.3 The Population III.2 IMF

The previous section showed that the characteristic temperature of star-forming gas in this channel is much smaller than for Population III.1 stars, with $T \sim T_{\rm CMB}$. Such effective cooling leads to Bonnor-Ebert masses of ∼10–50 $M_\odot$, depending on the physics of cooling (see §5.2.4). This likely limits the masses of these Population III.2 stars to be just a few tens of solar masses, considerably below the upper limits on Population III.1.

Numerical simulations[20] show that a small protostar (with $m_\star < 0.5\ M_\odot$) forms in the Population III.2 case, just as in the case without deuterium, and subsequent stages proceed similarly to that case as well. However, in the colder gas, fragmentation into smaller mass protostars is much more likely, and the protostars are very unlikely to grow to the ∼100 $M_\odot$ scales necessary to make radiative feedback relevant. Thus, it appears plausible that the Population III.2

IMF is skewed toward high-mass stars, but stars that still lie within the mass range observed in the nearby Universe.

This second-generation process may therefore produce a much different IMF than the first generation. However, we have seen that turbulence, chemical processes, and gravitational instability may cause even Population III.1 protostellar systems to fragment into clumps of comparable sizes. It remains to be seen how different these two formation channels really are.

5.4 Properties of the First Stars

We have so far examined the formation mechanisms of primordial stars; we now move on to discuss the stars themselves, and especially their radiative outputs and observable characteristics. Note that once the mass of a zero-metallicity star is set through its formation mechanism, its properties are independent of that mechanism. Thus, we refer to Population III stars without specifying their subgroup (III.1 or III.2) in this section.

If fragmentation is inefficient, Population III stars appear to grow orders of magnitude more massive than the Sun, probably ceasing accretion only when radiative feedback becomes important (§5.2.2). Primordial stars with $m_\star > 100\,M_\odot$ have an effective surface temperature $T_{\rm eff}$ approaching $\sim 10^5$ K, with only a weak dependence on their mass.[21] This temperature is $\sim$17 times higher than the surface temperature of the Sun, $\sim$5,800 K. These massive stars are held against their self-gravity by radiation pressure, which maintains their radiation field at the Eddington luminosity (see equation 5.27 and the discussion in §7.4), which is strictly proportional to their mass $m_\star$,

$$L_E = 1.3 \times 10^{40} \left(\frac{m_\star}{100\,M_\odot} \right) \text{ erg s}^{-1}, \tag{5.47}$$

and is six to seven orders of magnitude more luminous than the Sun, $L_\odot = 4 \times 10^{33}$ erg s^{-1}. Because of these characteristics, the total luminosity and color of a cluster of such stars simply depends on its *total* mass and not on the mass distribution of stars within it.

The radii of these stars $R_\star$ can be estimated by equating their luminosity to the emergent blackbody flux $\sigma_{\rm SB} T_{\rm eff}^4$ times their surface area $4\pi R_\star^2$ (where $\sigma_{\rm SB}$ is the Stefan-Boltzmann constant). Thus,

$$R_\star = \left(\frac{L_E}{4\pi \sigma_{\rm SB} T_{\rm eff}^4} \right)^{1/2} \approx 4.3 \times 10^{11} \text{ cm} \times \left(\frac{m_\star}{100\,M_\odot} \right)^{1/2}, \tag{5.48}$$

which is only approximately six times larger than the radius of the Sun, $R_\odot = 7 \times 10^{10}$ cm.

The high surface temperatures of the first stars made them ideal factories of ionizing photons: hydrogen requires an energy of 13.6 eV to liberate its electron, while helium requires 24.4 eV for the first electron and 54.6 eV for the second. These values are coincidentally near the characteristic energy of a photon

emitted by these very massive Population III stars but far above the characteristic energy of the Sun.

If indeed they were this massive, the first stars had a lifetime of a few million years, independent of their mass, because $L \propto m_\star$. During its lifetime, a very massive Population III star would have produced $\sim 10^5$ ionizing photons per proton incorporated in it; the precise efficiency depends on mass and the model parameters of the star, but only to within a factor of ~ 2 in the $m_\star = 10^2$–$10^3\ M_\odot$ range. This means that only a tiny fraction ($>10^{-5}$) of all the hydrogen in the Universe need be assembled into Population III stars for there to be sufficient photons to ionize the rest of the cosmic gas, a fact that may be important during the reionization process (see chapter 9). For comparison, Population II stars with a standard Salpeter IMF (equation 5.37) produce on average $\sim 4{,}000$ ionizing photons per proton in them.[22]

If fragmentation is permitted, the masses may be considerably smaller, $\sim$10–50 $M_\odot$, which is much larger than the characteristic mass today but still within the range of "normal" stars. In this case, the Population III stars will not be qualitatively different from their present-day analogs, although there are, of course, some differences in detail.

Evolutionary models of Population III stars are fairly well specified, with the primary uncertainty at the high-mass end being the degree of mass loss during stellar evolution. Figure 5.13 shows some example calculations. The solid lines show main-sequence evolutionary tracks for zero-metallicity stars without mass loss, while the short-dashed lines assume strong mass loss. Similar evolutionary tracks are shown for low-metallicity stars ($Z = 0.02\ Z_\odot$) with the dotted lines, and the zero-age main sequence for solar-metallicity stars is shown with the vertical solid line. Primordial stars tend to be hotter (or bluer) than their enriched counterparts (as well as slightly smaller). There are two reasons for this. First, the CNO cycle is inefficient (it is able to use only the small amount of carbon built up during the pre–main sequence phase). Thus, primordial stars have very hot cores. The lack of heavy elements also reduces the opacity of the outer layers. Together these factors imply hotter stellar surfaces.

For this reason, even lower-mass Population III stars are therefore somewhat more efficient at producing ionizing photons than Population II (or I) stars, but the difference is one of quantity rather than quality: the former emit $\sim$50% more ionizing photons per unit mass. The overall efficiency of producing ionizing photons therefore depends extremely sensitively on the IMF: only if very massive stars are indeed able to form will Population III stars be orders of magnitude more efficient than later generations of stars. Figure 5.14 illustrates this very important point: it shows the observed spectra of two Population III star clusters, one with purely very massive stars (solid line; in this case the spectrum is mostly independent of the mass distribution of the stars), and a standard Salpeter IMF (dotted line). For the same total stellar mass, the observable flux is larger by an order of magnitude for stars with masses $>100\ M_\odot$.

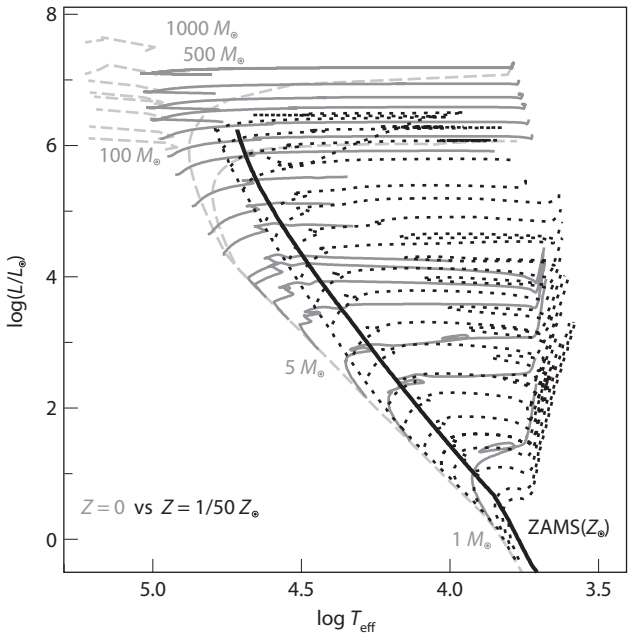


Figure 5.13 Main-sequence evolutionary tracks for Population III stars (solid lines: without mass loss; short-dashed lines: with strong mass loss) and $Z = 0.02\ Z_\odot$ stars (dotted lines). Isochrones at 2 and 4 Myr for the $Z = 0$ stars are also shown with the long-dashed lines. The zero-age main sequence for solar-metallicity stars is shown by the vertical solid line; note that Population III stars are significantly hotter (bluer) than their higher-metallicity counterparts. Schaerer, D., *Astron. & Astrophys.* **382**, 28 (2002).

5.4.1 Emission Lines: Signatures of Primordial Stars

The hotter temperatures and increased ionizing efficiencies of massive Population III stars imply that galaxies in which massive stars are prevalent will have some unique observational signatures. As the high-energy photons escape into the interstellar media of their host galaxies, many of those photons will encounter neutral hydrogen or helium and be absorbed. The ionized gas will then recombine, emitting one or more line photons as the atom returns to the ground state. The relative numbers of these line photons depend on the incident spectra and so can be used as diagnostics of the stellar IMF.

Let us define $\dot{Q}_{i,\star}$ as the rate at which a star of mass $m_\star$ produces photons capable of ionizing a species i. Because line emission is the result of absorbing these photons, we have for a line m

$$L_m = f_m h\nu_m (1 - f_{\rm esc})\dot{Q}_{i,\star}, \tag{5.49}$$

where $f_{\rm esc}$ is the fraction of photons that escape the galaxy without absorption, the choice of species i depends on the transition m, $h\nu_m$ is the energy of a

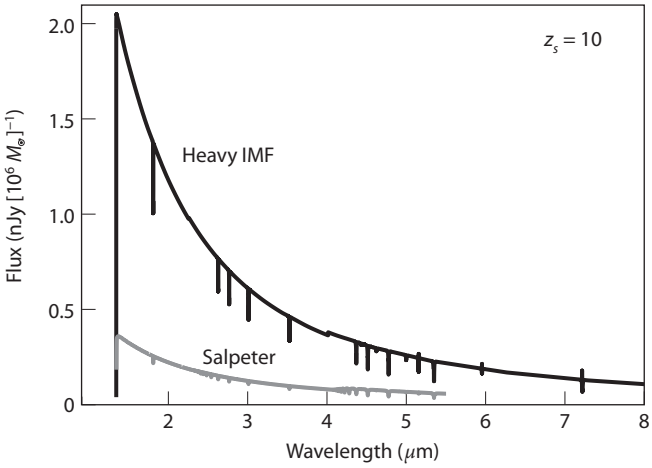


Figure 5.14 Comparison of the observed flux per unit frequency from a cluster of Population III stars at a redshift $z_s = 10$ for a Salpeter IMF (~~*dotted* line~~) and an IMF composed purely of very massive stars (~~*solid* line~~). The cutoff below an observed wavelength of 1216 Å $(1 + z_s) = 1.34\,\mu$m is due to hydrogen Lyman-α absorption in the IGM (the so-called Gunn-Peterson effect; see §4.3). Bromm, V., Kudritzki, R. P., & Loeb, A., *Astrophys. J.* **552**, 464 (2001); Salpeter curve from Tumlinson, J., & Shull, M. J., *Astrophys. J.* **528**, L65 (2000). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

light / dark / AA (dotted not showing up on revised figure)

by / AAS. /

photon emitted in transition m, and f_m describes how likely a recombination of the appropriate species is to produce a photon in this line.[iv] Because these last two factors depend only on atomic physics, the ratios of different lines provide the ratios of ionizing photons and hence a measure of the spectral hardness of the local stellar population, albeit modulated by the factor $(1 - f_{\rm esc})$, which could in principle also depend on frequency.

In fact, if all these ionizing photons are absorbed within the host galaxy, the hot, dense nebulae create substantial continuum emission as well, through free-free emission from the hot electrons, free-bound emission (by H I, He I, and He II) from the recombinations themselves, and the two-photon continuum of H I (generated when atoms recombine through the 2*s* level, which is metastable but eventually decays to the ground state by emitting two photons; see §12.2.2 for more on this process). This redistributes a large fraction of the energy contained in ionizing photons to lower frequencies and can substantially boost the brightness of the galaxies.

Figure 5.15 shows an example spectrum of a zero-age Population III star cluster in which the IMF contains high-mass stars but is not exclusively made up of them. The solid curve shows the spectrum including the reprocessing

[iv] This expression assumes that recombinations instantaneously follow ionizations. At the high densities characteristic of galaxies, this is a good approximation.

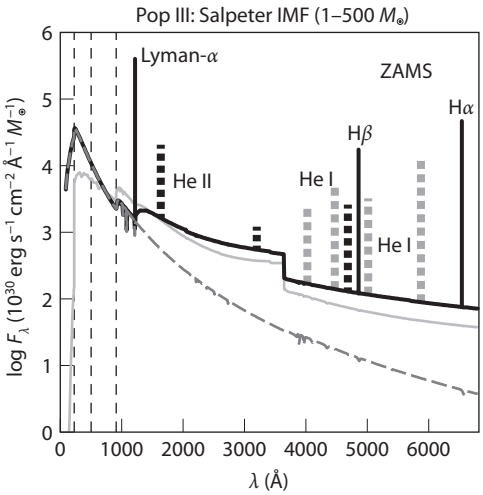


Figure 5.15 Spectral energy distribution of a cluster of Population III stars with a Salpeter IMF ranging from 1 to 500 $M_\odot$ (*solid line*), all of which have just entered the main sequence. Nebular reprocessing and recombination line emission are included assuming that $f_{esc} = 0$; emission lines are shown with solid, short-dashed, and long-dashed lines for H I, He I, and He II, respectively. The pure stellar continuum (neglecting nebular emission) is shown by the long-dashed line. The contrasting case of a Population II cluster with $Z = 0.02\ Z_\odot$ and a Salpeter IMF ranging from 1 to 150 $M_\odot$ is shown by the ~~dotted~~ line. The vertical dashed lines indicate the ionization potentials of H I, He I, and He II (from right to left). Schaerer, D., *A & A* **382**, 28 (2002).

light / AA (dotted not showing up on revised figure)

from nebulae and recombination lines; the long-dashed curve shows the stellar continua themselves. Because such a large fraction of the energy is originally invested in ionizing photons, this reprocessing enhances the rest-frame optical continuum by nearly an order of magnitude and creates very strong lines. Here, H I lines are shown with solid lines, He I with short-dashed lines, and He II with long-dashed lines.

The ~~dotted~~ curve shows the spectrum of a Population II cluster with $Z = 0.02\ Z_\odot$ and a Salpeter IMF ranging from 1 to 150 $M_\odot$ (normalized to the same total mass). The Population III case is somewhat brighter. More striking is the presence of the He II recombination lines at 1640, 3203, and 4868 Å, which appear because the highest-mass Population III stars are extremely hot and so release a substantial fraction of their energy (up to ~12%) above the He II ionization edge. In standard models, higher-metallicity (or lower-mass) stars produce almost no photons above this level, so these recombination lines are very interesting signatures of very massive Population III stars. (However, note that these lines *do* appear in some special stellar populations like Wolf-Rayet stars, and in some star-forming galaxies at lower redshifts.)

light / AA (dotted not showing up on revised figure)

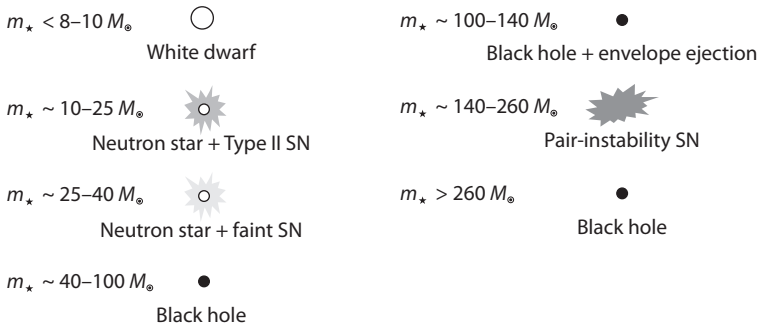


Figure 5.16 Likely fate of Population III stars based on their progenitor mass. Note that the mass ranges are only estimates and depend on both rotation and mass loss. Many of these supernovae may have unique observational signatures as well, such as gamma-ray bursts. See text for details.

However, because the highest-mass stars live for only a few million yeras, these He II recombination lines do not persist for long after the most recent burst of star formation. They are therefore not *necessary* signatures of zero-metallicity stars, even if they are convenient markers.

5.5 The End States of Population III Stars

The final result of the evolution of massive Population III stars is also important, both for observations and for the future evolution of the host halos. Several fates are possible, depending on the initial stellar mass.[23] The states themselves are easy to identify, but modeling supernovae is sufficiently difficult that the dividing lines between the different scenarios are uncertain. For example, rotation or strong magnetic fields can generally increase the mass thresholds identified next, though the degree of increase is not well quantified.

With these caveats in mind, the expected fates of Population III stars, and the rough divisions between them, are listed here and shown graphically in Figure 5.16.

- For masses below $m_\star < 8$–$10\ M_\odot$, stars end their lives as white dwarfs, just as present-day low-mass stars do. These stars can produce light elements during their asymptotic giant branch phases, but they do so ~~occurs~~ over much longer timescales than the <1 Gyr Hubble time at $z > 6$, so their fate is generally not considered important in understanding the histories of early galaxies.
- For masses $m_\star \sim 10$–$25\ M_\odot$, stars undergo Type II supernovae, leaving a neutron star behind. Especially at low metallicities, where the opacities are smaller, the hydrogen envelopes remain intact: these are the "normal" supernovae that are thought primarily responsible for the enrichment of very heavy elements in the nearby Universe.

PE

- For masses $m_\star \sim 25$–$40\ M_\odot$, stars undergo relatively weak Type II supernovae because much of the ^{56}Ni falls back onto the black hole remnant. As a result, these supernovae are likely quite faint and leave little iron behind.
- For masses $m_\star \sim 40$–$100\ M_\odot$, the stars collapse directly to a black hole *without* producing a supernova (and hence without enriching their surroundings), except through winds.
- For masses $m_\star \sim 100$–$140\ M_\odot$, the enormous core following helium burning heats up rapidly, leading to the production of electron–positron pairs as a result of collisions between atomic nuclei and energetic gammarays. This process reduces the thermal pressure inside the star's core, because some of that energy is lost in generating the rest mass of the pairs. This instability creates violent mass-ejecting pulsations, which can contain as much energy as a supernova (though is much fainter owing to the lack of radioactive elements). The entire hydrogen envelope of the star is likely ejected, relieving the instability and allowing the remainder of stellar evolution to proceed as for a lower-mass star, and the iron core eventually collapses directly to a black hole. These kinds of "explosions" do not significantly enrich their galaxies because only the light envelopes are ejected.
- For masses $m_\star \sim 140$–$260\ M_\odot$, stars are likely to explode as *pair-instability supernovae*.[24] A pair-instability supernova is triggered by the same instability already described, when part of the core's thermal energy is invested in the rest mass of electron–positron pairs. The pressure drop leads to a partial collapse and then greatly accelerated burning in a runaway thermonuclear explosion, which blows up the star without leaving a remnant. The kinetic energy released in the explosion can reach $\sim 10^{53}$ erg, 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, thus far there has been no conclusive evidence for their existence. 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 (§10.1.2). Their unusual explosion mechanism also imprints distinct nucleosynthetic signatures on pair-instability supernovae. They produce a near-solar distribution of elements from oxygen to nickel *except* with a large deficit of nuclei with odd charges, because weak interactions are unimportant throughout most of this mass range. They are also unable to make very heavy elements and eject no elements heavier than zinc.

  Furthermore, these pair-instability supernovae have very different observational properties than "normal" supernovae, as shown in Figure 5.17. The lines show three example light curves from modeling the explosion and shock breakout, while the curves with symbols show observed light curves of known supernovae. Pair-instability events for very massive stars occur over much longer timescales—of roughly a year—and are exceptionally luminous compared with more normal events. These long timescales are further exaggerated by cosmic time dilation as the photons travel

/ PE (insert space)

Close up space / PE

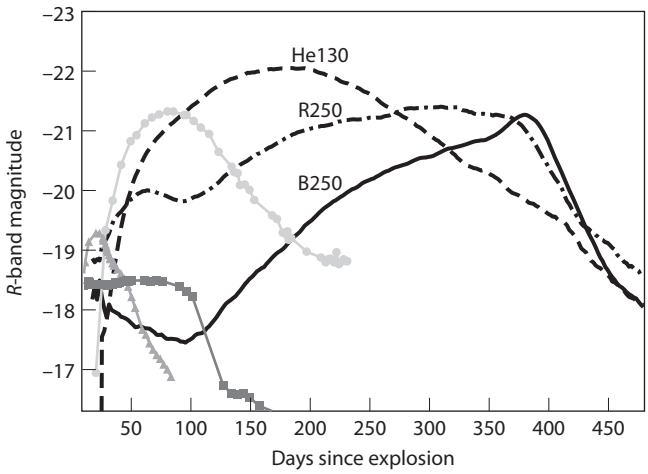


Figure 5.17 Example *R*-band light curves of three representative pair-instability supernovae (curves without symbols). The solid and dot-dashed curves show the explosions of blue and red supergiants with $m_\star = 250\ M_\odot$, respectively. The dashed curve shows the explosion of a bare helium core with $m_\star = 130\ M_\odot$ (this star is assumed to have lost its envelope through a wind and so is still subject to the instability). For comparison, the curves with symbols show three light curves of real supernovae. These are a normal Type Ia event (triangles), a normal Type IIP event (squares), and the overluminous core-collapse event SN 2006gy (circles). Kasen, D., Woosley, S. E., & Heger, A., *Astrophys. J.* **734**, 102 (2011). Reproduced with permission of the ~~American Astronomical Society~~.

by / AAS. /

through the expanding universe, making high-redshift events very intriguing from an observational point of view. The variations among the pair-instability models also show that careful observations can help constrain the progenitor mass, structure, and even metallicity.

- For masses $m_\star > 260\ M_\odot$, the helium cores instead collapse directly to black holes; nuclear burning of heavier elements is simply unable to halt the implosion triggered by exhaustion of more efficient fuel, and the entire star is swallowed up in the black hole, though possibly with a transient accretion disk and accompanying electromagnetic signature. Above this mass threshold, Population III stars therefore do *not* enrich their surroundings.

5.6 Gamma-Ray Bursts: The Brightest Explosions

Gamma-ray bursts (GRBs) were discovered in the late 1960s by the American *Vela* satellites, built to search for flashes of high-energy photons ("gamma rays") from Soviet nuclear weapon tests in space. The United States suspected that the

Soviets might attempt to conduct secret nuclear tests after signing the Nuclear Test Ban Treaty in 1963. On July 2, 1967, the *Vela* 4 and *Vela* 3 satellites detected a flash of gamma radiation unlike any known nuclear weapons signature. Uncertain of its meaning but not considering the matter particularly urgent, the team at the Los Alamos Laboratory, led by Ray Klebesadel, filed the data away for future investigation. As additional *Vela* satellites were launched with better instruments, the Los Alamos team continued to find unexplained GRBs in their data. By analyzing the different arrival times of the bursts as detected by different satellites, the team was able to estimate the sky positions of 16 bursts and definitively rule out either a terrestrial or solar origin. The discovery was declassified and published in 1973[25] under the title "Observations of Gamma-Ray Bursts of Cosmic Origin."

them / PE

The distance scale and nature of GRBs remained mysterious for more than two decades. Initially, astronomers favored a local origin for the bursts, associating 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 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 originated at cosmological distances, since the Universe is the only system that is truly isotropic around us. Nevertheless, the local origin remained more popular within the GRB community for 6 years, until February 1997, when the Italian-Dutch satellite *BeppoSAX* detected a gamma-ray burst (GRB 970228) and localized it to within a few arcminutes using its X-ray camera. With this prompt localization, ground-based telescopes were able to identify a fading counterpart in the optical band. Once the GRB afterglow faded, deep imaging revealed a faint, distant host galaxy at the location of the optical afterglow of the GRB. The association of a host galaxy at a cosmological distance for this burst and many subsequent ones revised the popular opinion in favor of associating GRBs with cosmological distances. This shift in popular view provides testimony to how a psychological bias in the scientific community can be overturned by hard scientific evidence.

A GRB afterglow is initially brightest at short photon wavelengths and then fades away at longer wavelengths, starting in the X-ray band (over timescales of minutes to hours), shifting to the UV and optical bands (over days), and ending in the IR and radio (over weeks and months).[v] Observers noticed that among the first detected afterglows, as the afterglow light curve faded, long-duration GRBs showed evidence for a supernova flare, indicating that they are also associated with core-collapse supernova events. The associated supernovae were classified as related to massive stars that have lost their hydrogen envelope in a wind. In addition, long-duration GRBs were found to be associated with star-forming regions where massive stars are born and explode only a million years afterward.

[v]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).

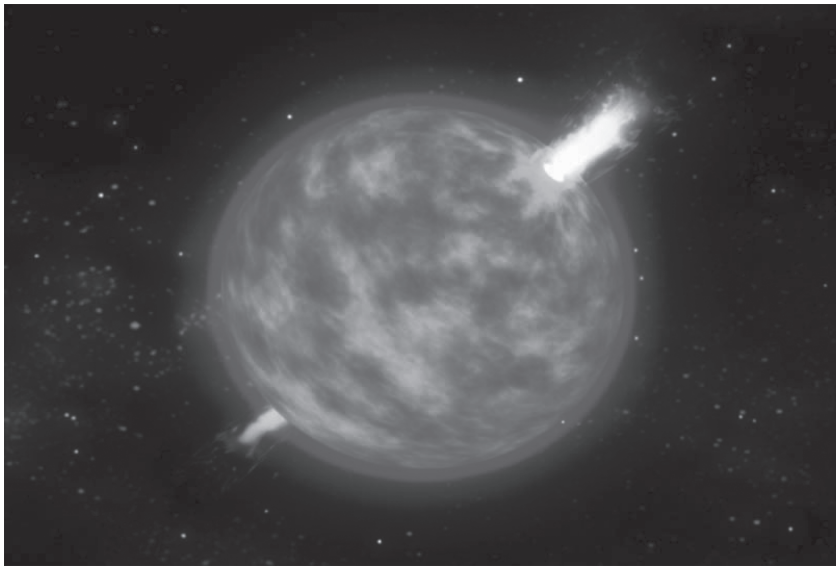


Figure 5.18 Illustration of a long-duration gamma-ray burst in the popular collapsar model (see *Color Plate 7* for a color version of this figure). The collapse of the core of a massive star (which lost its hydrogen envelope) to a black hole generates two opposite jets moving out at a speed close to that of light. The jets drill a hole in the star and shine brightly toward 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. NASA E/PO.

Courtesy of /

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 [26] (see Figure 5.18). 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 spirals into the black hole, two jets are produced at a speed close to that of light. So far, there is nothing spectacular about this setting, since we see scaled-up versions of such jets being formed around massive black holes in the centers of galaxies, as shown in Figure 7.3. However, when jets are generated in the core of a star, they make their way out by drilling a hole in the surrounding dense envelope. As soon as the head of a jet exits, the highly collimated stream of radiation emanating from it will appear as a gamma-ray flash to an observer who happens to line up with the jet axis. The subsequent afterglow results from the interaction between the jet and the ambient gas in the vicinity of the progenitor star. As the jet slows down by pushing against the ambient medium, the nonthermal 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 could 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[vi] at $z \sim 9.4$, representing the most distant source known, originating when the Universe was only ∼0.5 billion years old.[27]

It is unknown whether Population III stars produce long-duration GRBs. For that to happen, the angular momentum of the collapsing core mass M_c needs to be larger than $\sim 10 G M_c^2/c$, so that a stable disk will form outside the resulting black hole and collimate the jets. The rotation of the pre-GRB progenitor can be affected by mass exchange with a binary companion or mass loss through a wind. If the final mass of the black hole from a Population III progenitor is larger than usual, then the duration and total energy output of the associated GRB is expected to increase ($\propto m_\star$) relative to low-redshift GRBs. For additional details about observing GRBs, see §10.2.4.

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

Chapter Six

Stellar Feedback and Galaxy Formation

Chapter 5 described star formation in gas with a primordial composition, assuming that the star-forming region is completely isolated from its surroundings. However, as also discussed in the introduction to that chapter, such isolation cannot last long: those very same stars generate radiation fields that travel vast distances through the IGM. These photons can ionize the surrounding gas, drive winds or shocks through it, heat it, or photodissociate the H_2 or HD that is crucial for subsequent star formation. Moreover, supernovae or winds produced by these stars can also enrich the ambient gas with heavy elements. The second generation of stars that form in any given region is therefore influenced by its predecessors, and we must consider this influence. In this section, we study radiative, mechanical, and chemical feedback in the earliest gaseous clouds. While the feedback effects are sufficiently complex that a complete description of them is well beyond the capabilities of present-day computer simulations, the general principles that underlie them are well known. We therefore focus on these principles and then briefly sketch the global picture in §6.6.

Feedback is important in all galaxies, and many of the principles that we discuss in this chapter apply on a much wider scale than just the first stars and galaxies. We therefore present them in this larger context when appropriate, and return to their implications in chapters 7 and 8.

6.1 The Ultraviolet Background and H_2 Photodissociation

In the previous chapter we found that star formation in primordial gas depends crucially on molecular hydrogen to cool the cloud to densities high enough for stars to form. In this section we will consider what is likely to be the first important feedback process to affect this picture: how radiation from those very same stars can destroy that coolant and so make subsequent star formation even harder.

6.1.1 Lyman-Werner Photons and the Solomon Process

Molecular hydrogen (H_2) is fragile and can easily be photodissociated by photons with energies of 11.5–13.6 eV, to which the IGM is transparent even before it is ionized. The photodissociation occurs not through a single excitation step, as in the ionization of atoms but, instead, through a two-step process, first suggested by Phil Solomon in 1965.[1] In practice, this Solomon process is the only

way to photodissociate H_2 in interstellar (or intergalactic) space, because the photodissociation continuum of H_2 begins at 14.7 eV, while the photoionization continuum begins at 15.4 eV.[i] Both of these lie above the photoionization threshold of H I, so such photons would be absorbed by H I long before they encountered H_2.

The quantum-mechanical configuration of the electronic ground state of H_2 is denoted $X^1\Sigma_g^+$. Uppercase Greek letters denote the total electronic angular momentum of the system, projected onto the internuclear axis, with Σ, Π, and Δ having values of 0, 1, and 2 in units of $\hbar$. The left superscript (1 for the ground state) is $2S+1$, where $S=0$ or 1 is the total spin angular momentum. The right subscript (g or u) and superscript ($+$ or $-$, applicable only to Σ states) describe the symmetries of the configuration; this one asymptotes to two atoms with their electrons in the $1s$ state at large separations. The leftmost Roman letter describes the electronic states, with X being the lowest level, and the relevant upper states for our purposes labeled B and C (capitalized letters refer to singlets). Each of these electronic states is further split into a large number of sublevels by the quantized rotational and vibrational levels of the two nuclei, usually denoted by N and v. For example, the ground state has 14 vibrational levels, each nominally with an infinite number of rotational levels. Figure 6.1 shows a simplified level diagram for H_2, with these various splittings labeled.

The next two singlet states are $B^1\Sigma_g^+$ and $C^1\Pi_u$, which asymptote to two atoms with their electrons in the $1s$ and $2s$ or $2p$ states, respectively. These can decay to the ground state via permitted electric dipole transitions, the analog of the H I Lyman-α transition. However, for these molecules there are a large number of sub-transitions owing to the rotational and vibrational splittings. Thus, H_2 has two *bands* representing these transitions. The first band between the ground state and $B^1\Sigma_g^+$ is known as the *Lyman* band and consists of many densely packed lines beginning at 1108 Å (11.26 eV). The second band between the ground state and $C^1\Pi_u$ is known as the *Werner* band and begins at 1040 Å (12.3 eV).

Now consider the following sequence:

$$H_2(X^1\Sigma_g^+, v=0)+\gamma \rightarrow H_2(B^1\Sigma_g^+, v=v') \rightarrow H_2(X, v=v'')+\gamma \qquad (6.1)$$

Here v labels the vibrational energy level. Crucially, in electronic transitions there are no sharp selection rules for the vibrational continuum. Thus, the excited state's vibrational quantum number v' is not restricted to be small, and nor is the final value v''. It is therefore possible for the final state to lie in the vibrational continuum of the molecule ($v''>14$): in other words, to dissociate the molecule. A similar process also occurs for excitations and decays through the Werner band, as shown in Figure 6.1.

The rate at which these two steps occur depends on the crosssection for absorbing Lyman-Werner photons (for which the oscillator strengths are typically ∼1%) and the probability of decay into the dissociated continuum (typically

/ PE (insert space: cross section)

[i]These are far above the dissociation energy of H_2 (4.48 eV), because the direct transition is forbidden.

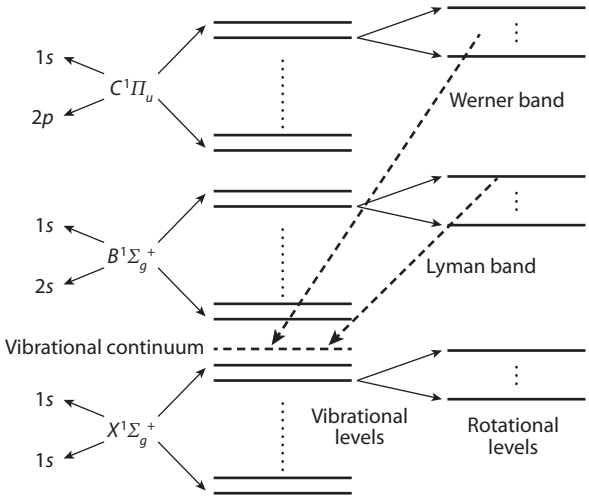


Figure 6.1 Energy level diagram for H_2. The left side of the diagram shows the energy level notation (such as $X^1\Sigma_g^+$), with the leftmost column showing the atomic hydrogen levels to which these molecular levels correspond at large separations. Each of these electronic levels splits into vibrational levels (shown in the third column), with a (dissociated) vibrational continuum above them. The electronic ground state contains 14 such levels. Each of these vibrational levels then splits into rotational levels, shown in the rightmost column. The Lyman (Werner) band corresponds to transitions from the first (second) excited state to the ground state; they are bands because of the many vibrational and rotational sublevels. Photodissociation occurs if one of these transitions leaves the molecule in the vibrational continuum.

$\sim$15%). The small vertical lines in Figure 6.2 show the energies of some of these transitions, where the initial configuration has $v = 0$ and $J = 0, 1$; the height of each line is $0.01 \times f_{\rm osc}$. The *average* cross-section for this process between 11.26 eV and 13.6 eV (averaged over 76 allowed lines) is $\sigma_{\rm LW} = 3.71 \times 10^{-18}$ cm^2.

6.1.2 The Suppression of H_2 Cooling

Once Lyman-Werner photons appear, we must include this photodissociation process in the chemistry of the primordial clouds.[2] The rate coefficient for photodissociation is

$$k_{\rm diss} = 1.38 \times 10^9 J_{\rm LW}\ {\rm s}^{-1}, \tag{6.2}$$

where $J_{\rm LW}$ is the specific intensity (in units of erg s^{-1} cm^{-2} Hz^{-1} sr^{-1}) in the Lyman-Werner band (specifically, here we have taken $h\nu = 12.87$ eV for concreteness, in the middle of the relevant energy range). It is convenient to normalize $J_{\rm LW} = 10^{-21} \times J_{\rm LW,21}$. The timescale for dissociation is therefore

$$t_{\rm diss} = k_{\rm diss}^{-1} \approx 3 \times 10^4 J_{\rm LW,21}^{-1}\ {\rm yr}, \tag{6.3}$$

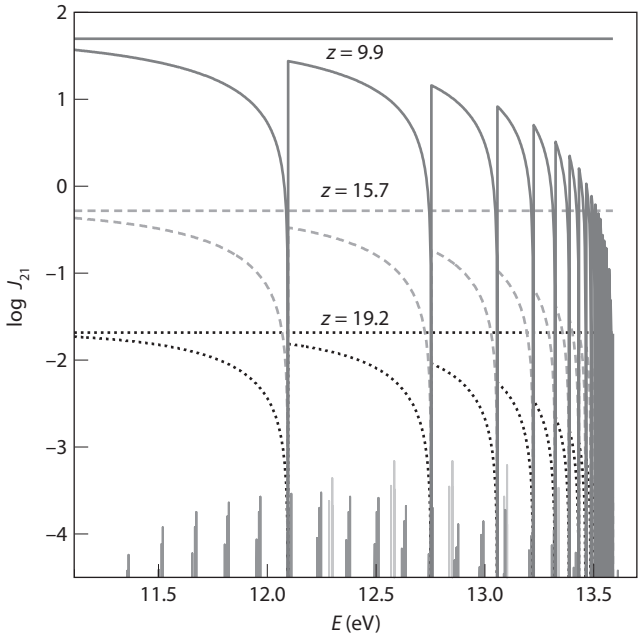


Figure 6.2 The "sawtooth" modulation of a uniform, spectrally flat radiation background in the Lyman-Werner frequency band when the IGM is still predominantly neutral (see *Color Plate 8* for a color version of this figure). The three curves are for $z = 19.2$, 15.7, and 9.2, from top to bottom; the horizontal lines show the unattenuated spectrum, while the curves with features show the effect of Lyman-series absorption. The vertical lines at the bottom of the figure show some of the Lyman-Werner transitions, with the height equal to 1% of the oscillator strength. Ahn, K. et al., *Astrophys. J.* **695**, 1430 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

which is very short compared to the relevant cosmological timescales. Thus, if the Lyman-Werner background approaches this fiducial value, we would expect it to destroy all the molecular hydrogen inside star-forming clouds.

In that case, if the radiation background and local gas properties remain constant on longer timescales, the H_2 fraction will approach an equilibrium in which the formation rate (approximately proportional to $\tilde{k}$ in equation 5.6) balances the dissociation rate,

$$f_{\mathrm{H_2,eq}} = \frac{\tilde{k}}{k_{\mathrm{diss}}} x_{\mathrm{H\ II}} n_{\mathrm{H}} \sim 4 \times 10^{-8} J_{\mathrm{LW},21}^{-1} \left(\frac{x_{\mathrm{H\ II}}}{2 \times 10^{-4}}\right) \left(\frac{1+z}{20}\right)^3 \left(\frac{\Delta}{200}\right), \quad (6.4)$$

where we have taken $T \sim 1{,}000$ K and a typical electron fraction before any cooling begins. This is far below the critical value required for H_2 cooling to be efficient (equation. 5.12), so *a substantial Lyman-Werner background suppresses molecular hydrogen cooling inside collapsed objects.*

The primary question is then whether a background of this amplitude can reasonably penetrate the clouds in which primordial stars may form. In the

next section we will examine whether a sufficiently strong background can be produced by the integrated stellar population, but before doing so we note that any such metagalactic radiation field must penetrate to the regions in which H_2 actually forms—that is, to the centers of virialized halos. Once H_2 cooling becomes important, these halos host large masses of the gas, and the outer layers of each halo can then *self-shield* the inner layers in which cooling actually occurs. If these outer layers are dense enough to maintain an equilibrium H_2 population that is optically thick in the Lyman-Werner bands, this self-shielding is significant. A convenient numerical approximation for the effects of self-shielding in a *static* medium is to take $k_{\rm diss} \rightarrow f_{\rm sh} k_{\rm diss}$, with[3]

$$f_{\rm sh} = \min\left[1, \left(\frac{N_{\rm H_2}}{10^{14}\ {\rm cm}^{-2}}\right)^{-0.75}\right], \tag{6.5}$$

where $N_{\rm H_2}$ is the column density of molecular hydrogen. The dependence at high column densities is steeper than expected from a naive curve-of-growth analysis (see §4.3.3 for a related discussion about H I Lyman-α) because of overlap within the various Lyman-Werner lines. (This estimate is not accurate at very high column densities, but those are rarely important in this context.)

Note, however, that self-shielding is more complex if the medium has velocity gradients, because then the lines are shifted by different amounts relative to their rest wavelengths in different parts of the cloud. This can considerably reduce the effectiveness of self-shielding and is a critical question in evaluating the importance of a Lyman-Werner background. In particular, virialized halos are continuosly accreting gas, which falls toward their center. The resulting velocity gradient helps keep the star-forming clouds at the centers of halos more optically thin than one would naively expect.

6.1.3 Photodissociation Feedback inside Star-Forming Halos

It is conceptually convenient to divide the Lyman-Werner photodissociating background into two simple cases: one in which light from a given star inside a collapsed halo acts upon gas inside the same halo, and a second in which a metagalactic radiation background affects halos from their exteriors. We will first consider the internal feedback case. We certainly expect that within some zone around an individual star, the Lyman-Werner background will dissociate enough H_2 to render further cooling inefficient, choking off later star formation. The critical question is the size of this zone compared to the halo.

We suppose that a star sits at the center of such a halo. To gauge the cumulative amount of H_2 destroyed around the star, we must compare the timescale for a star to photodissociate the halo's H_2 to its main-sequence lifetime. Very massive Population III stars produce $N_{\rm LW} \approx 3{,}400$ photons in the 11.2–13.6 eV range per baryon inside them; smaller stars produce them at about double that rate. If we assume that a fraction $f_{\rm LW,abs} \sim 0.01$ of these photons are absorbed by the Lyman-Werner bands (a reasonable approximation for the relevant column densities and expected line widths), and that about $f_{\rm LW,diss} \sim 0.15$ of these absorptions lead to dissociations (see §6.1.1), the total number of dissociations

from a star (or set of stars) with mass $m_\star$ is $\sim f_{\rm LW,abs} f_{\rm LW,diss} N_{\rm LW} m_\star / m_p$. Comparing this to the total number of H_2 molecules in a halo, $\sim f_{H_2} M_g / m_p$ (where M_g is the total gas mass), we find that the fraction of molecules expected to be photodissociated is

$$f_{\rm destroy} \sim 10^4 \left(\frac{f_{\rm LW,abs}}{0.01}\right)\left(\frac{f_{\rm LW,diss}}{0.15}\right)\left(\frac{N_{\rm LW}}{3400}\right)\left(\frac{f_{H_2(s)}}{3.5\times10^{-4}}\right)^{-1}\left(\frac{m_\star}{M_g}\right). \quad (6.6)$$

Italics / PE

Thus, provided that the star formation efficiency is not extremely small, the first generation of stars can easily photodissociate all of their halo's diffuse H_2, shutting down further cooling at least temporarily.

However, gas clumps already in the process of collapse may be dense enough to maintain their H_2 populations in the presence of this radiation background. The relevant question for clumps is whether the radiation field can dissociate the H_2 both before collapse completes (over $\sim t_{\rm dyn}$) and faster than the clump can form H_2 to replace it (equation 6.4). Analytic estimates show that clumps that have already passed the "loitering stage" (with $n_{\rm H} > 10^4$ cm^{-3}) are suppressed only very close to the source star.[4] Thus, the total rate of star formation within halos may depend on the degree to which clumps are synchronized across the entire halo: those collapsing at nearly the same time will be unaffected by the Lyman-Werner background, but the collapse of those that are delayed may be halted completely. In particular, gas clumps fragmenting from the same disk are very unlikely to be affected by each other's Lyman-Werner photons, because they are synchronized.

6.1.4 The Metagalactic Lyman-Werner Background

Because the IGM is mostly optically thin to photons in the Lyman-Werner bands (and the small amount of intergalactic H_2 is quickly dissociated as the first sources appear), a metagalactic radiation field will quickly build up in this energy range. If the background is intense enough, the rate at which H_2 is destroyed inside collapsed objects will exceed the rate at which such molecules form, preventing cooling in newly forming halos—and causing a strong *negative* feedback effect on star formation.

The magnitude of this feedback will depend on how these Lyman-Werner photons propagate through the IGM. In fact, the IGM is not perfectly optically thin to them, as absorption by the H I Lyman-series lines processes the background below the Lyman limit, causing the sawtooth shape shown in Figure 6.2. 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 by one of the (extremely optically thick) Lyman-series resonances along the way.[ii] Consider, for example, an energy of 11 eV at an observed redshift $z = 10$. Photons emitted just below the 12.1 eV Lyman-β line from $z = 11.1$ would be received at 11 eV at

[ii]The Lyman-α optical depth is given in equation (4.11), and higher Lyman-series transitions fall proportionally to the ratios of the oscillator strength times frequency.

$z = 10$. Thus, sources in the redshift interval 10–11.1 could be seen at 11 eV, but radiation emitted by sources at $z>11.1$ would have passed through the 12.1 eV Lyman-β energy at some intermediate redshift, and would have been absorbed.

It follows that 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 was above the nearest Lyman resonance: photons with energies just below a Lyman resonance would be contributed by only a small number of sources, while those just above one of these energies would be sourced out to relatively large distances. Most of the photons absorbed along the way would be reemitted either at Lyman-α or in the $2p \rightarrow 1s$ two-photon continuum and then redshift to lower energies. The result is a sawtooth spectrum for the UV background before reionization, with an enhancement below the Lyman-α energy due to reprocessing.

Quantitatively, the specific intensity at a frequency ν and redshift z is (see equation 4.43)

$$J_\nu(z) = \frac{c}{4\pi} \int dz' \frac{dt}{dz'} \left(\frac{1+z}{1+z'}\right)^3 \epsilon_{\nu'}(z') e^{-\tau(z)}, \tag{6.7}$$

where $\epsilon_{\nu'}(z')$ is the proper emissivity from sources at a redshift z' and a frequency $\nu' = \nu(1+z')/(1+z)$, and the factor $\tau(z)$ is the accumulated optical depth as the photon travels through the IGM. This is negligible so long as the photon stays between the Lyman-series lines, but it becomes very large whenever the photon crosses such a line. A simple but accurate approach is therefore to use a "screening approximation" in which the integral is truncated at a maximum redshift determined by the nearest Lyman line i (of frequency $\nu_i > \nu$) via

$$\frac{1+z_{\rm max}}{1+z} = \frac{\nu_i}{\nu}, \tag{6.8}$$

while the optical depth factor can otherwise be ignored.

Figure 6.2 shows this modulation in detail for a set of uniform emissivity sources with flat spectra at three different final redshifts (the normalizations are arbitrary; the horizontal lines show the spectra before attenuation by the Lyman series). As the frequency increases and the spacing between the Lyman-series lines decreases, the absorbing screens get closer together, and the total background decreases. Thus, the uppermost Lyman-Werner transitions experience a weaker background.

Unfortunately, the direct detection of the redshifted sawtooth spectrum as a remnant of the reionization epoch is not feasible owing to the much higher flux contributed by foreground sources at later cosmic times. However, a similar process does occur before He II is completely reionized at $z<3$, and the Lyman-series transitions of He II create a sawtooth spectrum in the far-UV. This spectrum may be indirectly detectable through its effects on metal-line absorbers, some of whose ionization potentials lie inside the sawtooth region

of the spectrum. This effect is barely visible in the higher-redshift panels of Figure 4.13.

Estimating the spectrum in more detail, and as a function of redshift, requires a model for the emissivity $\epsilon_\nu(z)$. Clearly, that will depend on the galaxy formation processes that we will examine over the next several chapters, but for a very simple estimate we can assume that the star formation efficiency $f_\star$ within halos is zero below a minimum halo mass $M_{\rm min}$ and constant above that mass (which could be set either by accretion—the filter mass—or efficient cooling to form stars). Then, we can write

$$\epsilon_\nu(z) = \frac{1}{4\pi} f_\star \frac{df_{\rm coll}}{dt} \frac{\bar{\rho}_b}{m_p} \epsilon_{\rm LW}(\nu), \tag{6.9}$$

where the first factor converts from total emissivity to emissivity per solid angle, and the last factor is the energy produced by the stars per frequency per baryon in the Lyman-Werner region. If we approximate the latter by $\epsilon_{\rm LW} \approx h\nu_{\rm LW} N_{\rm LW}/\Delta\nu_{\rm LW}$, equation (6.7) gives

$$J_{\nu,21} \sim 2.4 \left(\frac{N_{\rm LW}}{3400}\right) \left(\frac{f_\star}{0.1}\right) \left(\frac{\Delta f_{\rm coll}}{0.01}\right) \left(\frac{1+z}{10}\right)^3, \tag{6.10}$$

where $\Delta f_{\rm coll}$ is the fraction of gas that collapses onto star-forming halos over the redshift range $(z, z_{\rm max})$. Radiation backgrounds of this magnitude are easily large enough to strongly suppress H_2 cooling in just-virialized gas (see equation 6.4).

Figure 6.3 shows a more careful calculation of the background spectrum amplitude, though still in the context of a model with the star formation rate proportional to $df_{\rm coll}/dt$, and $f_\star = 0.1$. Here we show the *average* amplitude over the entire Lyman-Werner frequency interval—the sawtooth absorption typically reduces this from the emitted amplitude by about an order of magnitude. We show several different mass thresholds, increasing from the filter mass (top curve) to masses near the atomic cooling threshold (bottom curve). The amplitude increases rapidly with decreasing redshift because these halos are initially on the exponential tail of the mass function; the turnover at lower redshifts is where the corresponding halos are well below the cutoff in the mass function, so that the growth slows down. Equation (6.10) appears to provide a reasonable estimate of $J_{\rm LW,21}$.

The choice of $f_\star$ is highly uncertain in these models, so Figure 6.3 is only a very rough guide to expectations. If the first cluster of Population III.1 stars shuts down further star formation in a halo, then one might expect only a few hundred solar masses of stars to form inside each one. In that case, $f_\star = m_{\star,\rm tot}/M_g \sim 0.003 (m_{\star,\rm tot}/500\ M_\odot)(M_h/10^6\ M_\odot)^{-1}$, where M_h is the total halo mass. Fortunately, these curves are all strictly proportional to the star formation efficiency, so their amplitude can easily be rescaled.

This mean background is relatively easy to compute, but in reality the clustered halos that source the background induce inhomogeneities in it. Fortunately, at least in the standard structure formation model, these inhomogeneities are mild. Consider the lower edge of the Lyman-Werner band,

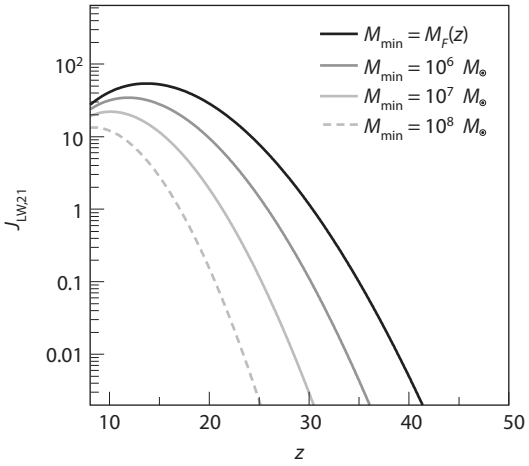


Figure 6.3 Evolution of the specific intensity of the metagalactic radiation field in the Lyman-Werner band at high redshifts. The ~~solid~~ lines show the amplitude of the radiation field over time, taking several different mass thresholds for star-forming halos: $M_{min} = M_F(z)$ (the filtering mass), 10^6, 10^7, and 10^8 $M_\odot$, from top to bottom. The curves assume $f_\star = 0.1$; all the curves are proportional to this value.

AA

with 11.2 eV. Photons redshift into this band out to the Lyman-β transition at 12.1 eV, which corresponds to a redshift of $\Delta z \sim 0.1(1+z)$, or about 100 Mpc. Each point therefore samples a huge volume of sources around it, which averages out the fluctuations (although the more closely spaced higher Lyman-series transitions weight the effective volume to more nearby sources). The Lyman-Werner background will therefore be nearly uniform except very close to individual sources or unless the halo population itself has fluctuations on ~100 Mpc scales,[iii] which may indeed be possible due to a strong source bias and the velocity offset between dark matter and baryonic material (see §2.1.2 and 3.2.2). In that case, the background may vary strongly, leading to substantial variations in the halos able to cool and form stars efficiently across very large scales.

6.1.5 External Feedback on H_2 inside Virialized Halos

With a model for the Lyman-Werner background in hand, it is now straightforward to gauge the metagalactic background's effects on H_2 cooling inside collapsing dark matter halos. As a simple estimate of the column density of a virialized halo, we assume a uniform density sphere at the typical virial

[iii] We will explicitly compute fluctuations in the background at the Lyman-α frequency based on a very similar calculation in §12.3.1.

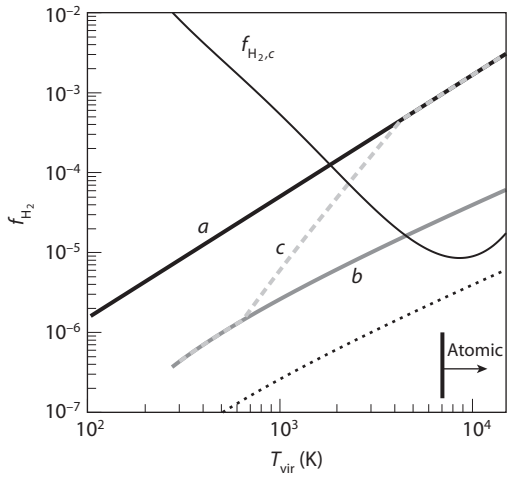


Figure 6.4 Schematic illustration of molecular hydrogen fraction as a function of virial temperature for halos inside a cosmological simulation at $z = 17$. The line at upper right shows $f_{H_2,c}$, the critical fraction required for efficient cooling. The dark solid line (marked *a*) shows $f_{H_2} \sim f_{H_2,s}$ in the absence of radiative feedback (see Figure 5.5). The light solid (labeled *b*) and short-dashed lines show f_{H_2} if self-shielding is neglected and $J_{LW,21} = 0.01$ or 0.1, respectively. The long-dashed line (labeled *c*) shows the same, but with self-shielding approximately included. Yoshida, N., et al., *Astrophys. J.* **592**, 645 (2003). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~ by / AAS. /

overdensity and with a radius r_{vir}. Then a halo of mass M_h has

$$N_{H_2} \sim 10^{17} \left(\frac{f_{H_2}}{3.5 \times 10^{-4}} \right) \left(\frac{M_h}{10^6\ M_\odot} \right)^{1/3} \left(\frac{1+z}{20} \right)^2 \ \mathrm{cm}^{-2}, \qquad (6.11)$$

where we have inserted the saturation value for the H_2 fraction from equation (5.10) as a fiducial estimate. In fact, simulations show that the effective column density is typically a few times smaller than this, since much of the gas in the outskirts of the halo remains optically thin, but the simulations confirm that the value provides a reasonable estimate for a stationary halo in which velocity gradients are insignificant (though that may not be a good approximation in reality).

This column density is well above the self-shielding threshold in equation (6.5), which implies that much of the halo will be shielded from the metagalactic background. Therefore, we write the effective background as $f_{sh} J_{LW,21}(z)$. We can then insert this radiation field into equation (6.4) to determine the H_2 fraction in the presence of feedback. Finally, comparison of this fraction with the critical value required for cooling, $f_{H_2,c}$ in equation (5.12), determines whether the halo is able to continue cooling and form stars.

Figure 6.4 provides a schematic illustration of these effects, based on fits to numerical simulations (cf. Fig. 5.5). The line at upper right shows $f_{H_2,c}$, the

critical fraction required for efficient cooling. The dark solid line (marked *a*) shows f_{H_2} in the absence of radiative feedback; this lies very near the saturation level $f_{H_2,s}$ of equation (5.10). The light solid line (marked *b*) shows f_{H_2} if self-shielding is neglected and $J_{LW,21} = 0.01$. This markedly reduces f_{H_2} and quantitatively matches the estimates described in this section. However, the dashed line labeled *c* shows the same, but with self-shielding approximately included (and ignoring velocities). Halos near the critical cooling threshold are already very optically thick, so in practice the radiation background has significantly less of an effect than naively expected.

Nevertheless, the growing Lyman-Werner background most likely "self-regulates" the earliest stages of star formation. Within each star-forming halo, the first few stars create a strong Lyman-Werner background and prevent any protostars not already far along in their collapse from proceeding. The same stars create a metagalactic background that reduces the efficiency of cooling in other, newly forming gas clouds, raising the mass threshold for star formation. But as the abundance and mass scale of dark matter halos increases, the larger gas clouds more effectively self-shield their inner regions, allowing the background to increase, which in turn raises the mass threshold, and so on. Eventually the Lyman-Werner background becomes so intense that star formation is possible only through atomic cooling in halos with $T_{vir} \sim 10^4$ K, for which photodissociation is unimportant. However, recall that these halos ionize their own gas at the virial shock and so likely form stars through the Population III.2 (deuterium-mediated) channel described in §5.3. This Lyman-Werner background may therefore regulate the transition from very high mass primordial stars to the lower-mass channel.

6.2 The X-ray Background: Positive Feedback

The radiative feedback on H_2 need not be entirely 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.[5] This effect could counteract the destructive role of H_2 photodissociation. Unlike UV photons, X-rays can penetrate huge distances across the Universe, even at high redshifts. The mean free path through the mean IGM density of an X-ray photon with energy E is

$$\lambda_X \approx 11\, \bar{x}_{H\,I}^{1/3} \left(\frac{1+z}{10}\right)^{-2} \left(\frac{E}{300\ \mathrm{eV}}\right)^{3} \ \mathrm{Mpc}; \qquad (6.12)$$

thus, photons with $E > 1.5[(1+z)/15]^{1/2} \bar{x}_{HI}^{1/3}$ keV propagate an entire Hubble length before interacting with the IGM. Similarly, they can penetrate large columns of dense neutral gas inside collapsed halos. Thus, an X-ray background may be pervasive at high redshifts.

Insert space between H I / PE

X-rays interact with primordial gas by ionizing either helium or hydrogen. The resulting free electron can gain a large kinetic energy (equal to the

difference between the photon energy and the ionization potential), which it then deposits as a mixture of heat, collisional ionization, and collisional excitation. Typically, a fraction $f_i \sim x_{\rm H\ I}/3$ of the energy is deposited in ionizing other atoms. Thus, a 1 keV photon can result in ∼25 free electrons (we discuss this secondary ionization process further in §9.8.2). Because these free electrons catalyze H_2 formation, X-rays can exert *positive* feedback on primordial star formation.

Insert space/ PE

An X-ray background seems almost inevitable at high redshifts,[6] with a number of possible sources: (1) Very massive Population III stars are hot enough for their blackbody spectra to extend into the soft X-ray regime. (2) Quasars or "miniquasars" must begin to form at very high redshifts to produce the extremely luminous quasars seen at $z \sim 6$ and likely have nonthermal spectra extending to very high energies. (3) Supernova blastwaves may accelerate fast electrons, which can in turn scatter CMB photons to X-ray energies. The associated cooling rate of relativistic electrons increases dramatically with redshift, since the CMB energy density scales as $u_{\rm CMB} \propto (1+z)^4$. (4) X-ray binaries, in which a massive black hole accretes gas from a companion, are often produced when a massive star explodes in a binary system; if massive stars are more abundant at high redshifts, then such binaries may be more common then. We will see later (§12.3.2) that these contributions to the X-ray background significantly affect the IGM temperature and ionization history, and they also present an important potential positive feedback mechanism for the first stars.

Simple scaling laws suggest, however, that this positive feedback will overcome the negative Lyman-Werner feedback only in unusual circumstances. Let us suppose that the electron fraction inside a cool cloud is in ionization equilibrium with an X-ray background. We will assume that the X-rays are sourced by the same population of galaxies as the UV background (though the sources themselves may differ, such as high-mass stars and the X-ray binaries they become after dying).

For an X-ray background amplitude J_X, ionization equilibrium implies $n_e \propto (J_X n_{\rm H})^{1/2}$, where we have ignored the temperature dependence of the recombination coefficient. Equation (6.4) therefore yields (with $x_{\rm H\ II} n_{\rm H} = n_e$)

$$f_{\rm H_2,eq} \propto n_{\rm H}^{1/2} J_X^{1/2} / J_{\rm LW}. \tag{6.13}$$

In other words, the equilibrium molecular fraction depends more weakly on the X-ray background than on the UV background. Assuming these are tied to the same underlying physical processes (i.e., are both ultimately driven by gas accretion onto halos and star or black hole formation), X-rays can make a substantial difference only when $J_{\rm LW}$ is still relatively modest. (Moreover, they matter at all only if the equilibrium electron fraction is larger than the value obtained from the usual chemistry described in §5.1.1.)

More detailed investigations have shown that if $J_X = \epsilon_X J_{\rm LW}$ at the H I ionization edge, X-rays exert mild positive feedback on dense gas clouds when $0.1 < \epsilon_X < 1$.[7] At smaller fluxes, the X-rays are relatively unimportant. At larger

fluxes, the heating generated by the X-rays counteracts the additional cooling, negating the boost to the free-electron fraction.[iv]

6.3 Radiative Feedback: Mechanical Effects

As discussed in §5.2.2, radiative feedback from the first stars may have been crucial for choking off accretion and setting their final mass scale. But the high-energy photons responsible for that process likely reached well outside the accretion disk, into the source halo and the surrounding IGM, once the star entered the main sequence. The same processes mentioned previously can dramatically affect these larger scales and subsequent star formation in a star's environment, because the radiation can influence the motion of the surrounding gas more than gravity does. In this section we consider some of the relevant processes in more detail.

6.3.1 The First H II Regions: Photoevaporation

The most dramatic effects result from the high luminosity of ionizing photons produced by the first stars. We discussed briefly in §5.2.2 how ionization fronts can have powerful effects on gas dynamics, and these effects extend far beyond the protostellar region once a star enters the main sequence. For example, consider an ionizing front expanding inside a gravitationally bound halo, where the baryon density declines with radius. For pedagogical purposes, we adopt a simple density profile:

$$n_{\rm H}(r) = \begin{cases} n_c & r < r_c, \\ n_c (r/r_c)^{-w} & r \geq r_c, \end{cases} \tag{6.14}$$

where w is a power-law index that encapsulates the steepness of the density run in the cloud's outskirts, and n_c and r_c are a core density and radius, respectively. Numerical simulations show that primordial gas clouds have $w \sim 2$–2.2 (see also §5.2.1). The average density of virialized (uncooled) gas inside dark matter halos at redshift z is $\sim 1\ {\rm cm}^{-3}[(1+z)/30]^3$, independent of halo mass.

The properties of the ionization front can be characterized with reference to the *Strömgren radius* R_s, the outer boundary of the H II region around the source out to which the total rate of recombinations is equal to the total rate of ionizations (see also §9.1). For a star producing ionizing photons at a rate $\dot{Q}_i$ in a constant-density medium, this radius is

$$R_s = \left(\frac{3\dot{Q}_i}{4\pi n_{\rm H}^2 \alpha_B} \right)^{1/3} \approx 150 \left(\frac{\dot{Q}_i}{10^{50}\ {\rm s}^{-1}} \right)^{1/3} \left(\frac{n_{\rm H}}{1\ {\rm cm}^{-3}} \right)^{-2/3} {\rm pc}, \tag{6.15}$$

where we have evaluated the recombination coefficient $\alpha_B = 2.6 \times 10^{-13}$ cm^{-3} s^{-1} at $\sim 10^4$ K (see equation 4.17). If the H II region reaches this size (or,

[iv] A similar negative feedback effect functions even at lower X-ray fluxes in the diffuse IGM, where X-ray heating creates an "entropy floor" that prevents gas from collapsing onto virialized objects. We discuss this effect in detail in §9.9.

alternatively, if within a fixed radius the density exceeds an equivalent threshold value), then the ionizing photons themselves are consumed within mostly ionized gas.

Before this time, the front is slowed only by the rate at which photons can ionize the medium. During the early fast-expansion phase, we refer to the ionization front as *R-type* (also see §5.2.2). However, once this expansion velocity slows down to near the sound speed, the gas is able to react to its new thermodynamic properties. The Strömgren radius provides a simple maximal estimate for when this transition occurs, because at that point the expansion has nearly zero velocity. In more detail, the ionization front slows to become *D-type* when its expansion speed falls to roughly twice the isothermal sound speed of the ionized medium, $2c_i$. At that point, the increased temperature (and hence pressure) within the front drives a shock into the surrounding medium. The front then propagates outward at roughly the speed of sound. The H II region can therefore expand only through hydrodynamic processes, and the ionization front is said to be *trapped*.

In the density profile given by equation (6.14), some algebra shows that the Strömgren radius is $R_w \equiv g(w)R_{s,c}$, where $R_{s,c}$ is evaluated with equation (6.15) using the core density and[8]

$$g(w) = \begin{cases} \left[\dfrac{3-2w}{3} + \dfrac{2w}{3}\left(\dfrac{r_c}{R_{s,c}}\right)^3\right]^{1/(3-2w)} \left(\dfrac{R_{s,c}}{r_c}\right)^{2w/(3-2w)} & w \neq 3/2, \\ \left(\dfrac{r_c}{R_{s,c}}\right) \exp\left\{\dfrac{1}{3}\left[\left(\dfrac{R_{s,c}}{r_c}\right)^3 - 1\right]\right\} & w = 3/2. \end{cases} \tag{6.16}$$

The front's speed depends on how far it extends: it can *accelerate* at $r > r_c$ if the density profile is steep enough. In particular, if $w > 3/2$, the total recombination rate ($\propto n_{\rm H}^2$ times the volume) does not appreciably increase as the front's radius grows, allowing the front itself to escape to infinity. To see this, it is straightforward to estimate the velocity at which the ionization front expands before the R_w limit is met:

$$U_{\rm i-f} = \frac{U_c}{(R_{s,c}/r_c)^3 - 1} u(w), \tag{6.17}$$

where U_c is the typical speed within the uniform-density core,

$$U_c \approx 90 \left(\frac{n_{\rm H}}{10^3\ {\rm cm}^{-3}}\right)\left(\frac{r_c}{10^{17}\ {\rm cm}}\right)\left[\left(\frac{R_{s,c}}{r_c}\right)^3 - 1\right]\ {\rm km\ s}^{-1}, \tag{6.18}$$

and

$$u(w) = \begin{cases} \left(\dfrac{r_c}{R}\right)^{2-w}\left[\left(\dfrac{R_{s,c}}{r_c}\right)^3 + \dfrac{2w}{3-2w} - \dfrac{3(R/r_c)^{3-2w}}{3-2w}\right] & w \neq 3/2, \\ \left(\dfrac{r_c}{R}\right)^{1/2}\left[\left(\dfrac{R_{s,c}}{r_c}\right)^3 - 1 - 3\ln(R/r_c)\right] & w = 3/2. \end{cases} \tag{6.19}$$

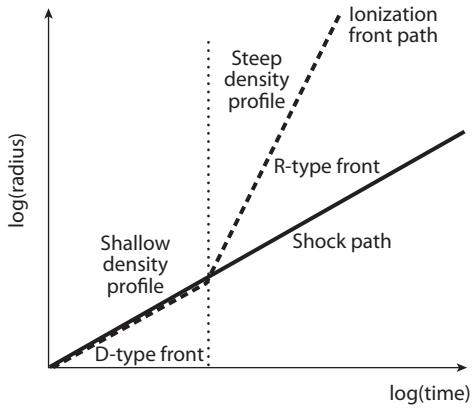


Figure 6.5 Diagram of an ionization front propagating through a cosmological halo. The dashed and solid lines show the locations of the ionization front and shock, respectively. After an initial R-type phase (not shown), recombinations in the high-density core trap the front, making it D-type, in which a shock slightly leads the ionization front. As the density falls through the halo, the recombination rate also falls, eventually freeing the front to expand much faster than the sound speed. The shock is left behind and lags the front, often eventually transforming into a simple pressure wave.

The ionization front remains R-type all the way to infinity if

$$w > w_{\rm trap} = \frac{3}{2}\left[1 - \left(\frac{r_c}{R_{s,c}}\right)^3\right]^{-1}, \qquad (6.20)$$

or $w > 3/2$ for ionization fronts able to reach well outside the core before striking the Strömgren limit.

The front shifts to D-type, driving a shock into the surrounding gas, if $w < w_{\rm trap}$. This allows the ionization front to grow (slowly), even though it has nominally reached its Strömgren limit, because the hydrodynamic motions of the gas decrease the average density behind the shock. In a typical halo, the density profile steepens as one moves outward, usually with $w > w_{\rm trap}$ in the outskirts. Therefore, the front eventually reaches a point where it is no longer trapped. At this time it reverts to R-type and expands rapidly, with no immediate hydrodynamic effect on gas outside of the H II region itself. Numerical simulations show that this transition point is well approximated by the Strömgren radius of the initial density profile,[9] using equation (6.15) with the average density set to its value inside R_s. Figure 6.5 shows a diagram of this evolution.

These have different spacing conventions.

However, within the H II region, the gas rapidly accelerates outward. The temperature structure of the cloud is set by photoheating: each ionization leaves the residual electron with some extra energy that depends on the spectrum of the ionizing source (see §9.9 for more details on this process), typically with $T \sim 10^4$ K. The pressure profile is set by the density profile—which also has not had time to adjust to its new state. A strong pressure gradient therefore

develops, producing an acceleration (again, with $\rho \propto r^{-w}$)

$$a = \frac{1}{\rho}\frac{dp}{dr} \sim \frac{wc_i^2}{r}, \tag{6.21}$$

which is strongest in the center of the halo. A pressure wave therefore develops, pushing the gas ahead of it out of the halo—this regime is often referred to as the *champagne phase*. Behind the wave, the gas has roughly constant density and hence reaches pressure equilibrium; ahead of it the gas is still in its original configuration.

The characteristic speed of this wave is a few times the sound speed of the ionized gas, $c_i \sim \sqrt{kT/m_p} \sim 10(T/10^4\ \mathrm{K})^{1/2}\ \mathrm{km\ s^{-1}}$. In comparison, the escape speed from a dark matter halo is roughly

add italics h^{1/3}

$$v_{\rm esc}(M) \approx \sqrt{2} V_c(r_{\rm vir}) = 33.0 \left[\frac{\Omega_m}{\Omega_m(z)}\frac{\Delta_c}{18\pi^2}\right]^{1/6} \left(\frac{M_h}{10^8 M_\odot}\right)^{1/3} \left(\frac{1+z}{10}\right)^{1/2} \mathrm{km\ s^{-1}}, \tag{6.22}$$

where we have used equation (3.31) and assumed an isothermal density profile truncated at the virial radius $r_{\rm vir}$, for simplicity. Thus, the gas inside the H II region becomes strongly unbound and flows outward for halos of a sufficiently low mass. The ionization front slows down only when it reaches a region with a shallower density gradient in the IGM, allowing it to return to the Strömgren limit. But by this point the bound gas has already escaped.

Numerical simulations of this *photoevaporation* process show that in the limit of a smooth, spherical halo, the radiation pressure from a single very massive Population III.1 star can evacuate the gas from an entire halo of mass $\sim 10^6\ M_\odot$.[10] Figure 6.6 shows an example from a detailed numerical calculation of a single $200\ M_\odot$ star. Clockwise from top left, the panels show the ionized fraction, the temperature, the (outward) velocity, and the density profile at a sequance of times. In the last panel, the dashed line shows the density required to enforce the Strömgren criterion in equation (6.15); if the density exceeds this value, the ionization front will be limited by recombinations and be D-type. Clearly, the high core density will trap the front, from which it will emerge when it reaches $\sim$1 pc. In the upper panels, the large jump in the ionization front location from 82 to 95 kyr involves the transition from D-type to R-type. The large outward gas velocities, at two to three times the sound speed of the gas, are nearly ten times as large as the escape speed from the minihalo (2–3 $\mathrm{km\ s^{-1}}$).

Thus, the first stars can easily empty their halos of gas, decreasing the local baryon fraction to just a few percent. However, the picture is less clear if the gas filling the source halo is clumpy, containing other collapsing cores (or in nearby halos also en route to forming their own stars). If these neighboring clumps have modest densities, they, too, will be completely evaporated. However, if their central densities are sufficiently high, $n_{\rm H} > 2{,}000\ \mathrm{cm^{-3}}$, the core will remain neutral via self-shielding, and the radiation will have little effect. In this case, the collapse will continue until new stars are formed. Indeed, the passage of an ionization front (and accompanying shock) through surviving cores may actually aid collapse and encourage further star formation. Whether more

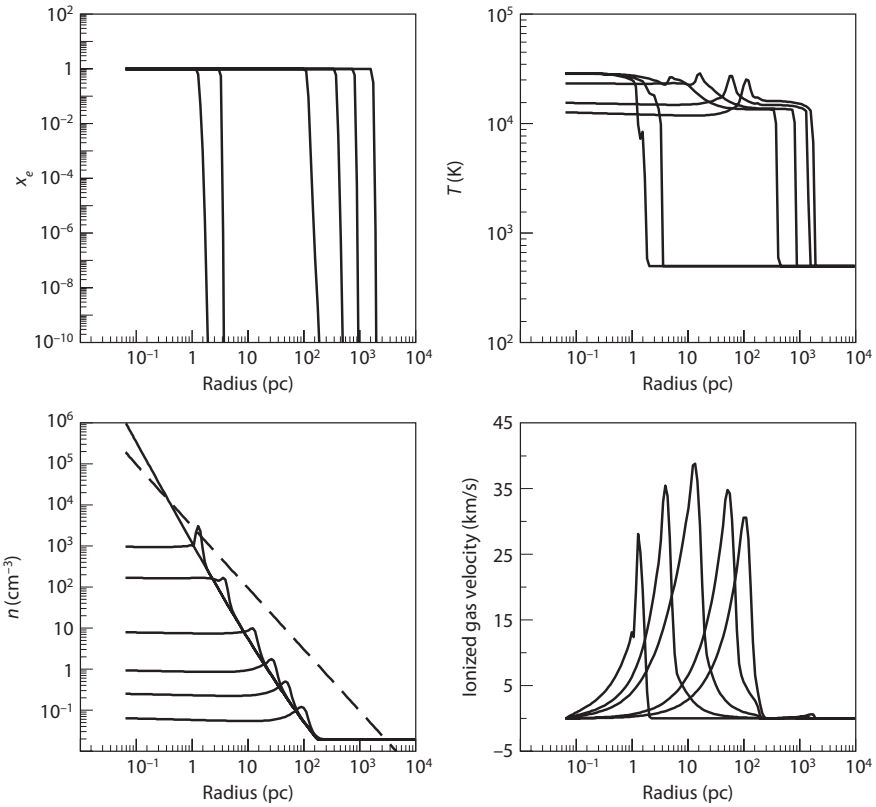


Figure 6.6 Evolution of a cosmological halo as an ionization front propagates through it. The simulation takes a single 200 $M_\odot$ star at $z = 18.2$ in a halo of total mass 7×10^5 $M_\odot$. Clockwise from top left, the panels show the ionized fraction, the temperature, the (outward) velocity, and the density profile. Each panel shows snapshots at 63, 82, 95, 127, 317, and 2200 kyr (left to right in all panels except bottom left, where they are top to bottom). In the bottom left, the dashed line shows the minimum density required to trap the ionization front. Whalen, D., et al., *Astrophys. J.* **610**, 14 (2004). Reproduced ~~with~~ by / permission of the ~~American Astronomical Society.~~ AAS. /

than one star can form in a low-mass halo (or in a halo with nearby neighbors) thus crucially depends on the degree of synchronization of clump formation.

As an example of the complex implications of the photoevaporative flow, consider the shocked gas that lies ahead of the front during its D-type phase. This shocked region is partially ionized by high-energy photons, and so its ionized fraction is typically appreciable ($>10^{-3}$; see §9.8.2 for more on this process). The extra free electrons catalyze the formation of H_2, potentially making self-shielding effective. The cooling induced by H_2 can trigger thin shell instabilities that quickly develop into new star-forming clumps.

The long-term effects of this radiation pressure are also not obvious and depend on the details of the halo's neighborhood. Although the gas has a very

high velocity as it leaves the halo, it can still be reincorporated into the halo (or into one of its nearby neighbors) through hierarchical structure formation. Numerical simulations show that this fallback can take ∼100 million years, a substantial fraction of the age of the Universe at these high redshifts.[11] This could lead to a long delay in later star formation or accretion onto any remnant black holes. The preionization would also change the mode of any future star formation to Population III.2 stars, possibly with a somewhat lower mass scale than the first generation of Population III.1 stars.

6.3.2 Radiation Pressure from Lyman-α Photons

Interestingly, the radiation can also exert a substantial force on the neutral gas surrounding the H II region.[12] The Lyman-α photons, generated primarily by recombinations within the H II regions, scatter off the neutral gas outside those regions, imparting their net outward momentum and driving the gas away from the central source. We can gauge the possible dynamical effect of these photons by comparing the gravitational binding energy, $E_B \sim (\Omega_b/\Omega_m) G M^2/r_{\rm vir}$ (see equation 3.33), with the energy in the radiation field, $E_\alpha = L_\alpha \times t_{\rm trap}$, where L_α is the line luminosity of the H II region, and $t_{\rm trap}$ is the typical timescale over which Lyman-α photons are trapped inside the cloud. Numerical calculations of line transfer suggest that $t_{\rm trap} \sim 15 t_{\rm light}$,[13] where $t_{\rm light} = r_{\rm vir}/c$ is the light travel time across the halo (see further discussion of this complex problem in §11.1.1). The condition $E_\alpha > E_B$ requires that

$$L_\alpha > L_{\alpha,{\rm crit}} \sim 10^{40} \left(\frac{M}{10^6\ M_\odot}\right)^{4/3} \left(\frac{1+z}{30}\right)^2 \left(\frac{15 t_{\rm light}}{t_{\rm trap}}\right) {\rm erg\ s^{-1}}. \quad (6.23)$$

Note that approximately two-thirds of recombinations produce a Lyman-α photon, so this translates to a direct constraint on the ionizing luminosity; the fiducial luminosity shown here corresponds to only ∼500 $M_\odot$ (per $M \sim 10^6 M_\odot$) in very massive Population III.1 stars, assuming that the H II region reaches its Strömgren limit.

For a nearly isotropic radiation field (a valid approximation in this case because of the large number of scatterings each Lyman-α photon experiences), the acceleration induced by Lyman-α radiation pressure may be written as

$$a_{{\rm Ly}\alpha} = \frac{1}{3\rho} \frac{dU_\alpha}{dr}, \quad (6.24)$$

where $\rho = m_p n_{\rm H}$, and U_α is the energy density of the Lyman-α photons. If the gas were optically thin, then U_α would be $L_\alpha/(4\pi r^2 c)$, but the scattering process traps Lyman-α photons near the source and steepens the $1/r^2$ scaling. The total impulse $a_{{\rm Ly}\alpha} \Delta t$ therefore depends on the total Lyman-α fluence of the source, which in turn is dictated by the number of ionizing photons produced by the stars.

The simple solution described in §11.6, for scattering around a point source in a uniform IGM expanding at the Hubble flow, has $U \propto r^{-2/3}$ at moderate distances from the source. In a more realistic calculation, the H II region

surrounding the central star, the infall region surrounding the halo, and the details of Lyman-α scattering must be taken into account, but this simple solution provides a reasonable gauge of the importance of Lyman-α radiation pressure. Assuming very massive Population III.1 stars, the corresponding final velocity of an atom at a distance r from the central source is

$$v_\alpha \sim 6 \left(\frac{1\ \mathrm{kpc}}{r}\right)^{10/3} \left(\frac{15}{1+z}\right)^3 \left(\frac{f_\star}{10^{-3}} \frac{M}{10^6\ M_\odot}\right)\ \mathrm{km\ s^{-1}}. \tag{6.25}$$

While the final velocity is small, the escape speed at the virial radius $r_{\rm vir} = 0.2$ kpc of a $10^6\ M_\odot$ halo at $z = 14$ is $\sim$6 km s^{-1}. Thus, Lyman-α scattering through the neutral gas *outside* any H II region can eject the gas from the vicinity of the source halo, also slowing down accretion.

This same effect can also operate in larger galaxies later in the history of structure formation, many of which are observed to have substantial Lyman-α fluxes. However, numerical simulations show that the effects are modest unless the galaxy also drives a wind that creates a neutral "supershell" that can multiply the radiation force through repeated scatterings, largely because these galaxies are able to ionize such a large region around them that the near region, where the force is strongest, is still ionized and cannot trap the photons.

6.4 Galactic Superwinds and Mechanical Feedback

6.4.1 Star Formation and Wind Energetics

As stars live and die, they inject large amounts of energy into their surroundings, through a number of channels. First, while they are luminous, their radiation couples to the interstellar medium as UV photons scatter off atoms and dust grains (which are usually coupled to the neutral or ionized gas through collisions and magnetic fields). Just as in the Lyman-α scattering case described previously, the pressure of the radiation field can therefore eject gas from the galaxy. Second, in the late stages of stellar evolution, many stars drive powerful winds into the ISM, and supernova explosions when massive stars die inject $E \sim 10^{51}$ erg of energy into the ISM, typically accelerating $\sim$10 $M_\odot$ of material per explosion to $\sim 3 \times 10^3$ km s^{-1}.

The energy and momentum flux from these mechanical interactions can unbind the gas from the host halo. These outflows, when they span a large fraction of a galaxy, are known as *superwinds*. By removing gas from the galaxy, these mechanisms choke off the fuel supply for further star formation and may ultimately be responsible for regulating the pace of star formation over time. A clear understanding of the role of feedback is therefore essential to understanding not only the first galaxies but their more massive descendants.

We begin with some plausibility arguments showing that winds are likely to be important for the small galaxies most common at high redshifts. We first ask the question, How much star formation is necessary to unbind the gas inside a virialized halo? The total binding energy of a halo with mass M is

given by equation (3.33), but for the *gas* we must multiply this energy by the mass fraction in gas ($f_g \sim \Omega_b/\Omega_m$). Moreover, we have already seen that to form stars the gas must collapse to high densities. To describe this situation simply, we assume that the gas is confined to a region $<\lambda r_{\rm vir}$ (see §8.4), where the spin parameter $\lambda \sim 0.05$ is set by the angular momentum of the halo. Thus, the gas binding energy is $E_{\rm b,g} \sim (f_g/\lambda) G M_h^2/r_{\rm vir}$.

Meanwhile, the energy injected by supernovae is $E_{\rm SN} \sim f_\star f_g M_h \omega_{\rm SN}$, where $f_\star$ is the fraction of gas that is turned into stars, and $\omega_{\rm SN}$ is the supernova energy input per unit mass of star formation. Typical supernova models and "normal" Population II IMFs yield $\omega_{\rm SN} \sim 10^{49}$ erg $M_\odot^{-1}$. However, we expect that some fraction of this energy will be radiated away as the hot, dense supernova remnant plows through the galaxy into the IGM around it. We assume that a fraction ξ of the total energy is available for mechanically removing gas from the galaxy. Then, the energy input by supernovae exceeds the binding energy of the gas if the star formation efficiency surpasses a critical value

$$f_\star > f_{\star,E} \sim 0.01 \left(\frac{0.05}{\xi\lambda}\right) \left(\frac{M_h}{10^8\ M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \left(\frac{\omega_{\rm SN}}{10^{49}\ {\rm erg}\ M_\odot^{-1}}\right)^{-1}. \quad (6.26)$$

Even if the supernova remnants do lose their thermal energy, they will still inject a great deal of momentum into the ISM. If this momentum is large enough, it can carry the gas outside the halo without the "push" from the thermal energy inside each remnant (i.e., feedback can be much more effective than suggested by equation (6.26) if $\xi \ll 1$).[14] The rate at which momentum is injected by supernovae, $dP_{\rm SN}/dt$, is

$$\frac{dP_{\rm SN}}{dt} \sim 2 \times 10^{33} \left(\frac{\omega'_{\rm SN}}{300\ {\rm km\ s^{-1}}}\right) \left(\frac{\dot{M}_\star}{M_\odot\ {\rm yr^{-1}}}\right)\ {\rm g\ cm\ s^{-2}}, \quad (6.27)$$

where $\omega'_{\rm SN}$ is the rate of momentum injection from supernovae per unit mass of stars; the fiducial value takes one explosion per 100 $M_\odot$ of stars, each accelerating 10 $M_\odot$ of material to 3×10^3 km s^{-1}.

Meanwhile, the rate at which stellar *radiation* injects momentum is $dP_{\rm rad}/dt \sim L_\star/c$, where $L_\star$ is the stellar luminosity that couples to the ISM gas.[v] We write it in terms of the rest energy as $L_\star = \epsilon \dot{M}_\star c^2$, where $\dot{M}_\star$ is the rate at which mass is processed into stars, and $\epsilon_3 \equiv (\epsilon/10^{-3}) \sim 1$ for typical IMFs. Then, $dP_{\rm rad}/dt \sim \epsilon_3 dP_{\rm SN}/dt$, which indicates that both sources of momentum are likely important in launching winds. For convenience, we write $dP/dt \equiv L/c$ for the total rate of momentum input, which defines an effective luminosity L that includes stellar radiation and supernovae (as well as any other processes, like winds from evolved stars).

The acceleration equation for a parcel of gas with velocity v and position r is[15]

$$\frac{dv}{dt} = -\frac{GM(r)}{r^2} + \frac{L}{cM_g(r)}, \quad (6.28)$$

[v] We assume here that the dust and gas are marginally optically thick to the radiation, so that they efficiently absorb the momentum flux.

Insert new text here:
If the scattering (but not absorption) optical depth is larger than unity, then the momentum injection rate increases in proportion to the optical depth, as the photons scatter back and forth multiple times.

where $M(r)$ is the halo mass, and $M_g(r)$ is the gas mass enclosed within a radius r. For a simple estimate, let us assume that the halo is a singular isothermal sphere, with $M(r) = 2\sigma^2 r/G$ and σ the velocity dispersion, and that the gas traces the dark matter. Then, we can rewrite equation (6.28) as

$$\frac{dv}{dt} = \frac{2\sigma^2}{r}\left(\frac{L}{L_M} - 1\right), \tag{6.29}$$

where

$$L_M = \frac{4 f_g c}{G}\sigma^4. \tag{6.30}$$

Clearly, L_M represents the minimum luminosity for the net force on the gas parcel to act outward, and hence it is the minimum luminosity required to launch a wind. If we further assume a constant star formation efficiency $f_\star$ to convert gas into stars over a dynamical time $t_{\rm dyn}{\sim}r_{\rm vir}/\sigma$, this minimum luminosity translates into a minimum star formation efficiency $f_{\star,p}$:

$$f_\star > f_{\star,p}{\sim}0.05(\omega'_{\rm SN,300} + \epsilon_3)^{-1}\left(\frac{M_h}{10^8\ M_\odot}\right)^{1/3}\left(\frac{1+z}{10}\right)^{1/2}, \tag{6.31}$$

where $\omega'_{\rm SN,300} = \omega'_{\rm SN}/(300\ {\rm km\ s^{-1}})$, and the $(\omega'_{\rm SN,300} + \epsilon_3)$ factor accounts for both supernovae and radiation.

If we compare equations (6.26) and (6.31), it is clear that for the small halos in which the first stars formed, the energy reservoir was likely much more effective than the raw momentum, provided it was not lost through radiative cooling. It is also clear that the required star formation rate in these halos was very small, fundamentally because the energy available in stars scales with M_h (assuming a constant $f_\star$), while the binding energy scales as M_h^2.

However, at higher masses the excess energy becomes less important, and the momentum injection condition becomes more stringent when

$$M{>}M_p{\sim}10^{10}(\omega'_{\rm SN,300} + \epsilon_3)^{-3}\left(\frac{0.05}{\lambda\xi}\right)^{-3}\left(\frac{1+z}{10}\right)^{3/2} M_\odot. \tag{6.32}$$

Insert - (minus sign) in superscript / PE

Nevertheless, for the momentum to lift gas out of the halo, star formation must proceed very quickly—turning a substantial fraction of the gas into stars over a single dynamical time. Such rates do appear in rapidly star-forming galaxies at lower redshifts, but those systems are relatively rare.

These two types of winds, *energy driven* and *momentum driven*, are likely to have very different characters. The condition that $E_{\rm SN}{>}E_{\rm b,g}$ does not place any restrictions on the rate at which *mass* is ejected from the galaxy; in fact, numerical simulations of star-forming disk galaxies typically show that the energy is "blown out" along low–column density channels (perpendicular to the disk), carrying away only a fraction of the galaxy's mass.[16] However, the momentum must maintain its initial direction and sweep up any gas it encounters as it propagates outward, carrying with it a significant fraction of the galaxy's gas. The asymptotic velocity v_∞ of a momentum-driven wind is typically just a few

times the escape speed of the halo.[vi] Momentum conservation then demands that the mass loss rate in the wind be

$$\dot{M}_w = \frac{dp/dt}{v_\infty} \sim \dot{M}_\star \left[\frac{300(\omega'_{\rm SN,300} + \epsilon_3)\ {\rm km\ s^{-1}}}{v_\infty} \right], \tag{6.33}$$

comparable to the star formation rate for reasonably large halos.

6.4.2 Expanding Blast Waves: Simple Solutions

To better understand the dynamics of these winds, we review here some simple models for expanding blast waves following point explosions. Although oversimplified, these analytic scalings provide useful insight into the more complex problem of winds inside and outside galaxies.

First, consider a point explosion with energy E in a static, cold (or pressureless) medium of mass density ρ. The explosion drives a shock into the surrounding gas. Simple dimensional arguments show that the shock radius must depend on ρ, E, and time t through the form

$$R_{\rm sh} = K_{\rm STV}(Et^2/\rho)^{1/5}, \tag{6.34}$$

where $K_{\rm STV}$ is a constant.

It is easy to show from energy conservation that $K_{\rm STV} \sim 1$. The total mass that is swept up by the shock is $\sim(4\pi/3)\rho R_{\rm sh}^3$. Because a supersonic shock forms in the ambient medium, the postshock gas velocity must be subsonic in the frame of the shock. Thus, most of the bulk velocity of the material is from the shock itself, and the net fluid speed is $\sim U_{\rm sh} = (2/5)R_{\rm sh}/t$. The kinetic energy of the swept-up material is therefore $\sim(4\pi/3)\rho R_{\rm sh}^3 \times U_{\rm sh}^2/2 \sim 0.3 R_{\rm sh}^5/t^2$. There is also, of course, the thermal energy stored in the hot gas behind the shock, but this is typically comparable to the kinetic energy of the shock, because the gas flows subsonically past the shock. Energy conservation implies that $E = \kappa\rho R_{\rm sh}^5/t^2$, where κ is a constant of the order of unity that accounts for summing the kinetic and internal energies. By comparison with equation (6.34), we see that $K_{\rm STV} = \kappa^{-1/5}$, which we expect to be very close to unity. In fact, this problem can be solved analytically and gives the exact value of $K_{\rm STV} = 1.17$ for a pure monatomic gas with an adiabatic index $\gamma = 5/3$. The solution is known as a *Sedov-Taylor–von Neumann blast wave*, after the three physicists who derived it independently at the dawn of the nuclear age. Since there is no characteristic timescale or length scale in the setup of a point explosion, the hydrodynamic equations admit a *self-similar* solution in which the hydrodynamic variables of the gas (pressure, density, and velocity) depend only on the combination $r/R_{\rm sh}(t)$ instead of depending separately on r and t.

The Sedov-Taylor–von Neumann solution imposes three restrictions on the blast wave. First, it requires that the mass of the material behind the shock

[vi] This can be seen, for example, by integrating equation (6.29) under the assumption that L is constant to obtain $v(r)$. Writing $(dv/dt) = (dr/dt)/(dv/dr) = v(dv/dr) = d[(\frac{1}{2})v^2]/dr$ and integrating both sides over $r > r_0$, we obtain $v(r) = 2\sigma \times [(L/L_M - 1)\ln(r/r_0) + v^2(r_0)/4\sigma^2]^{1/2}$.

be much greater than the explosion ejecta themselves. In an earlier phase, the ejecta expand ballistically, encountering negligible resistance from the ambient medium. Second, it requires a strong shock, or that the ejecta velocity greatly exceed the sound speed of the ambient medium. Finally, it requires that all the explosion energy is contained either in the kinetic energy or thermal energy of the shocked gas. In fact, the strong shock jump conditions require that the density just behind the shock be $(\gamma + 1)/(\gamma - 1)$ times that of the ambient medium (which is a factor of 4 if $\gamma = 5/3$) and decrease rapidly inward. This overdense shell cools radiatively; once a substantial fraction of the energy is lost, the energy conservation condition no longer applies, and the character of the solution changes. In particular, as the gas in the shell cools, its density must increase to maintain pressure equilibrium with the interior of the blast wave, and so a dense shell develops at the leading edge of the blast wave.

This second phase is known as a *pressure-driven snowplow*, because the low-density interior of the gas remains hot (and hence has a finite pressure p pushing outward on the shell). In this phase, the shell sweeps up gas as it expands, increasing its mass at a rate $\dot{M}_s = 4\pi R_{\rm sh}^2 \rho U_{\rm sh}$. Meanwhile, so long as the hot interior does not cool, the internal pressure obeys the adiabatic condition pV^γ = constant and pushes the shell outward with a force $4\pi R^2 p$. The equation of motion for the shell is then

$$\ddot{R}_{\rm sh} + \frac{3\dot{R}_{\rm sh}^2}{R_{\rm sh}} = \frac{3p_i}{\rho R_{\rm sh}} \left(\frac{R_i}{R_{\rm sh}} \right)^{3\gamma}, \tag{6.35}$$

where p_i is the internal pressure as this phase begins when $R_{\rm sh} = R_i$. For $p_i \neq 0$ and $\gamma = 5/3$, this equation requires that $R_{\rm sh} \propto t^{2/7}$, which is slightly slower than in the "adiabatic" Sedov-Taylor–von Neumann phase.[vii]

The pressure-driven snowplow phase ends when either one of two conditions is fulfilled. First, if the hot bubble interior can cool radiatively, it loses the pressure support. Second, if the interior pressure approaches the pressure of the ambient medium, there will be no net driving force. In either case, $p_i \rightarrow 0$ in equation (6.35). Then, $R_{\rm sh} \propto t^{1/4}$, which follows strictly from momentum conservation, $(4\pi/3)\rho R_{\rm sh}^3 (dR_{\rm sh}/dt) =$ constant. This final phase is therefore known as a *momentum-conserving snowplow*. Obviously, it is the proper solution for the momentum-driven winds described in the previous section.

So far we have assumed that the blast wave propagates into a uniform medium. While this describes the ISM of normal galaxies reasonably well (at least on average), the gas making the first stars did not settle into disklike configurations; instead, those stars were surrounded by uniform-density cores inside roughly power-law envelopes, with $\rho \approx \rho_0 (R_0/R)^\alpha$. Dimensional arguments similar to the preceding ones then show that

$$R_{\rm sh} = K_{\rm iso} \left(\frac{Et^2}{\rho_0 R_0^\alpha} \right)^{1/(5-\alpha)}. \tag{6.36}$$

[vii] Here the term *adiabatic* is standard in the literature. Note that it refers to a blast wave that conserves energy without radiative losses; obviously, the shock itself still increases the entropy of the gas.

W / PE

Table 6.1 Blast waves in Static Media. Each line shows a single phase in its evolution. The first column gives the abbreviated name (FE for free expansion, STvN for Sedov-Taylor–von Neumann, PDS for pressure-driven snowplow, and MCS for momentum-conserving snowplow); the second, the most important condition for the solution to apply; the third, the time dependence in a uniform medium; and the last, the time dependence in a medium where $\rho \propto r^{-\alpha}$.

Name	Condition	Growth rate (uniform)	Growth rate (power-law)
FE	Ejecta mass dominates	$R_{\rm sh} \propto t$	$R_{\rm sh} \propto t$
STvN	No shell cooling	$R_{\rm sh} \propto t^{2/5}$	$R_{\rm sh} \propto t^{2/(5-\alpha)}$
PDS	No interior cooling	$R_{\rm sh} \propto t^{2/7}$	$R_{\rm sh} \propto t^{2/(7-\alpha)}$
MCS	Cooled remnant	$R_{\rm sh} \propto t^{1/4}$	$R_{\rm sh} \propto t^{1/(4-\alpha)}$

In particular, for an isothermal density profile $\alpha = 2$, close to the envelopes of the first stars, $R_{\rm sh} \propto t^{2/3}$. The blast wave propagates faster in this case because the declining ambient density presents considerably less drag.

Similarly, it is straightforward to modify the equation of motion for the snowplow shell: since $\dot{M}_s = 4\pi R^2 \rho(R) U_{\rm sh}$, the momentum equation reads

$$\ddot{R}_{\rm sh} + \frac{(3-\alpha)\dot{R}_{\rm sh}^2}{R_{\rm sh}} = \frac{(3-\alpha)p_i}{\rho R_{\rm sh}} \left(\frac{R_i}{R_{\rm sh}}\right)^{3\gamma}. \tag{6.37}$$

For $\gamma = 5/3$, this equation admits of the solution $R_{\rm sh} \propto t^{2/(7-\alpha)}$ when the pressure is important, and $R_{\rm sh} \propto t^{1/(4-\alpha)}$ during the momentum-conserving snowplow phase. Again, specializing to an isothermal density profile, $R_{\rm sh} \propto t^{2/5}$ and $R_{\rm sh} \propto t^{1/2}$ in these two phases.

Table 6.1 briefly summarizes these different stages for quick reference.

6.4.3 Supernovae in the First Star-Forming Halos

The first supernovae occurred in the halos described in chapter 5. Although the basic properties of these halos are well-understood, the mass spectrum of stars and efficiency of star formation are highly uncertain, as they depend on the complex fragmentation process, the degree of synchronization of the resulting protostellar clumps, and the dynamical impact of the surrounding H II region.

For these reasons, the overall impact of the first supernovae on their host halos is difficult to assess. Nevertheless, numerical simulations have begun to explore these events and their implications for subsequent star formation, at least in some simple cases. Figure 6.7 provides an example, showing a simulated supernova explosion of a single star at $z \approx 20$ in a modest-sized halo. We examine this result in some detail because it illustrates much of the important physics of high-redshift supernovae.

The color scale shows the gas temperature. The large, roughly spherical region filling most of the box in all four panels is the H II region; its internal structure is a result of the filamentary cosmic web surrounding the halo. By

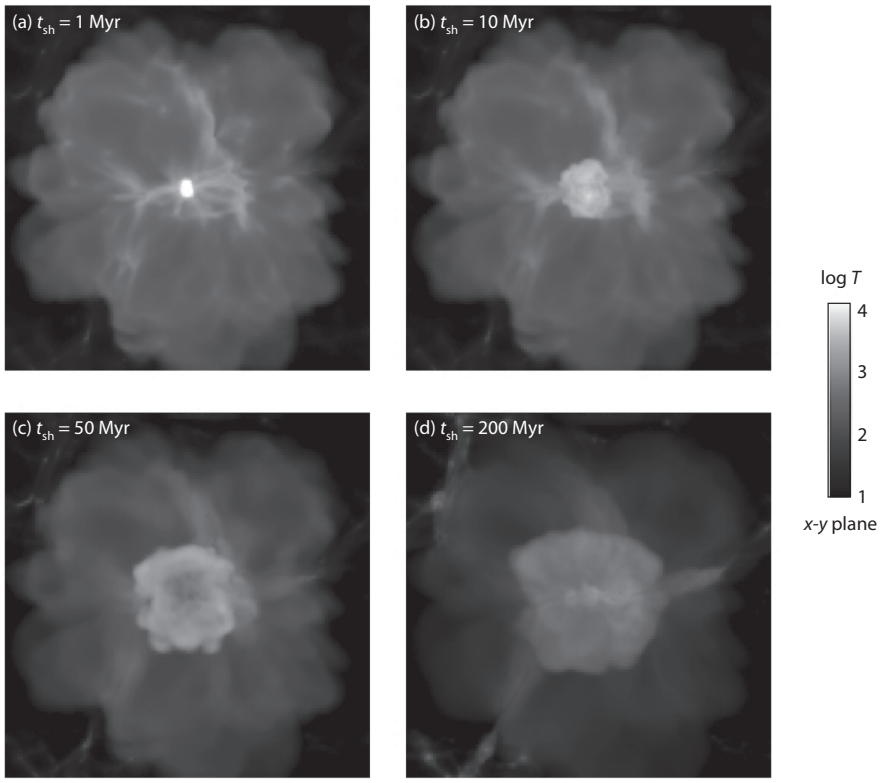


Figure 6.7 Temperature maps from a numerical simulation of a supernova explosion (see *Color Plate 9* for a color version of this figure). The supernova of a 200 $M_\odot$ star is set off at $z \approx 20$ in a halo with $M = 5 \times 10^5\ M_\odot$ and $r_{vir} \approx 100$ pc. The snapshots are 1, 10, 50, and 200 million years after the explosion. In the first panel on the top left, the supernova is the central hot region; the star's H II region fills most of the box (fading with time as the gas recombines and cools). The supernova remnant expands over the four panels, gradually becoming more anisotropic as it encompasses the filamentary structure surrounding the halo. Greif, T., et al., *Astrophys. J.* **670**, 1 (2007). Reproduced ~~with~~ permission of the ~~American Astronomical Society~~. by / AAS. /

the time of the star's death (2 Myr after its formation), it has photoevaporated the gas inside $\sim r_{vir}/2$, reducing its density to $n_H \sim 0.5$ cm^{-3}. Meanwhile, the escaping photons ionize a large region around the halo, initially heating it and causing pressure-driven expansion of the remnant into the low-density, cool IGM surrounding it.

The supernova then expands into this ionized environment. Figure 6.7 shows snapshots of its evolution, while Figure 6.8 presents the evolutionary phases of its (spherically averaged) radius. The four major phases of the expansion

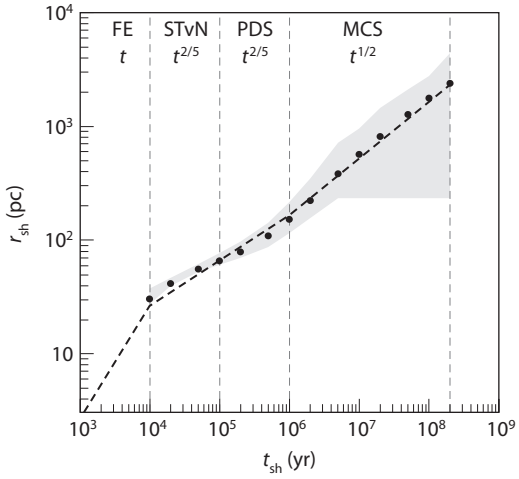


Figure 6.8 Evolution of the simulated supernova explosion described in Fig. 6.7. The black dots indicate the spherically averaged mass-weighted shock radius, while the dashed line shows the analytic estimate using the models of §6.4.2. The different phases in the evolution of the remnant are labeled: FE for free expansion (not resolved by the simulation), STvN for Sedov-Taylor–von Neumann phase, PDS for pressure-driven snowplow, and MCS for momentum-conserving snowplow. The shaded gray region shows the radial dispersion of the shock, which increases dramatically once the shock leaves its host halo owing to anisotropies in the cosmic web. Greif, T., et al. *Astrophys. J.* **670**, 1 (2007). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~ by / AAS. /

are marked in the latter figure. The explosion here, which is assumed to completely blow apart the star via a pair-instability supernova, carries a substantial mass $M_{\rm ej}$ in ejecta. Until the swept-up mass dominates the explosion, it expands freely (FE in Figure 6.8). The simulation does not follow this short phase explicitly; instead, it initializes the calculation at the end of this phase.

After that point, the Sedov-Taylor–von Neumann phase begins (labeled STvN). The blast wave initially propagates through a roughly constant density interior (the remnant gas after photoevaporation), so $R_{\rm sh} \propto t^{2/5}$. Once the remnant reaches $\sim r_{\rm vir}/2$ (at $t \sim 10^5$ yr), it catches up to the photoevaporation shock, and the character of its surroundings change. However, at just about this time the gas in the dense shell accumulating behind the shock is able to cool. Several processes allow cooling: atomic (and molecular) line radiation, bremsstrahlung, and inverse Compton scattering of CMB photons. Ignoring any possible chemical enrichment from the supernova itself, the atomic cooling rates are shown in Figure 5.1. Because these are driven by collisions, their rate scales as n^2. This mechanism is thus particularly important in dense gas, where it dominates over the other processes within the remnant shell. The cooling time is therefore $t_{\rm cool} \sim n_{\rm H} kT/\Lambda \sim 10^5$ yr, where the initial temperature is $T \sim 10^6$ K.

Thus, at about the same time the remnant reaches the photoevaporation shock, the shell gas cools, and the blast wave transitions to a pressure-driven snowplow solution (PDS in the figure). Now it propagates through the roughly isothermal spherical profile of the unperturbed halo, so $R_{\rm sh} \propto t^{2/5}$—coincidentally the same dependence as in the previous phase. This phase continues until either (1) the low-density interior gas is able to cool or (2) the postshock pressure reaches equilibrium with the ambient medium. At the very low densities characteristic of the remnant's interior, atomic cooling is inefficient. However, the cooling time due to inverse Compton scattering is independent of density (see equation 2.36),

$$t_{\rm cool} = 8 \left(\frac{20}{1+z} \right)^4 \ {\rm Myr}, \tag{6.38}$$

which puts an upper limit on the duration of this phase.

The postshock pressure reaches a value $p_{\rm sh} \sim \rho U_{\rm sh}^2 \sim p_{\rm H\ II}$, where $p_{\rm H\ II}$ is the pressure inside the H II region, after only $\sim 10^5$ yr. (This is easy to show using the analytic scalings of the previous section.) Thus, the blast wave transitions to its final phase, the momentum-conserving snowplow (labeled MCS in Figure 6.8), at $\sim 10^5$ yr. At the beginning of this phase, the density profile is still roughly isothermal, so $R_{\rm sh} \propto t^{1/2}$; in the simulation $R_{\rm sh}$ maintains this scaling even after passing into the IGM (see §6.5.2 for a discussion of solutions in this limit).

The net effect of this single supernova is to completely disrupt the gas in the host halo, expelling much of it (∼95%) and forcing the rest to high temperatures and low densities, where star formation is inefficient. The lack of star formation will persist until the high-entropy gas can be reincorporated through hierarchical buildup of higher-mass halos. Supernovae may therefore have been efficient in quenching star formation within the first star-forming halos.

However, as in so many other aspects of feedback, there are a number of subtleties to this simple picture, some of which may actually *promote* further star formation, including the following:

- First, the supernova itself is a source of heavy elements. As many of these elements are much more efficient low-temperature coolants than H or H_2, their presence could promote future star formation, particularly in combination with some of the mechanisms mentioned later. The primary uncertainty is the degree of mixing of the enriched material with the ambient medium, which is likely driven by instabilities in the shocked layers. We discuss the physics of this change in star formation mode in more detail in §6.5.
- Second, if the host halo remains largely neutral, the remnant will plow through much denser gas, even approaching $\sim 10^7$ cm^{-3}. Bremsstrahlung (free-free) cooling in such dense environments is extremely fast, and the supernova loses its thermal energy long before escaping the halo. Nevertheless, the impulse provided by the explosion can efficiently stir up the gas, possibly triggering further fragmentation as shells collide and most

likely dispersing heavy elements throughout the halo. This can happen, for example, if the characteristic mass scale of Population III.1 stars is only $\sim$10 $M_\odot$, and they form in massive halos, so that the explosion energy is lower than the gravitational binding energy of the halo gas.

- The blast wave itself may have very different effects on dense clumps (either inside the host galaxy or in nearby minihalos) than on the diffuse gas we have discussed. In particular, the shock compresses the gas, which increases the density, speeding up the later stages of collapse, provided the gas can efficiently cool (which is a precondition for star formation). Furthermore, the ram pressure of the shock is likely not able to move entire clumps along with the flow. The resulting configuration—a fluid stream flowing by a stationary cloud—can be unstable to Kelvin-Helmholtz modes. If so, the resulting mixing may allow metals to penetrate the outer layers of the pristine minihalo gas, triggering a change in the mode of star formation.
- Finally, the dense shell that accumulates behind the leading shock can itself be unstable and fragment through gravitational or cooling instabilities into protostellar clumps. The condition for such fragmentation is similar to the classical Jeans instability: collapse occurs when the self-gravity of the shell operates faster than its restoring pressure forces, which occurs on scales $> c_s/\sqrt{G\rho}$. For a given ambient density, the shell therefore eventually becomes unstable once the sound speed (or temperature) falls far enough, which of course requires efficient radiative cooling. In the case described in Figures 6.7 and 6.8, no fragmentation occurred because the low-density ambient medium both increased the dynamical time and inhibited molecule formation, maintaining relatively high temperatures. However, fragmentation can be much more efficient if the blast wave propagates through a denser neutral medium. In this case, the shell can trigger a second generation of stars. Because these stars form out of ionized gas (either from a preexisting H II region or one produced by the passing shock), they will be similar to Population III.2 stars discussed in §5.3, with lower characteristic masses.

6.5 Metal Enrichment and the Transition to Population II Star Formation

We have seen that the very first stars formed under conditions that were much simpler than the highly complex birthplaces of stars in present-day molecular clouds. As soon as the first stars appeared, however, the situation became more complicated owing to their feedback on the environment. In particular, supernova explosions dispersed the heavy elements produced in the interiors of the first generation of stars into the surrounding gas. Atomic and molecular cooling became much more efficient after the addition of these metals.

Early metal enrichment and dispersal by the primordial supernovae described in the previous section triggered a change in the fundamental mode

of star formation, because heavy elements can radiatively cool the gas much more efficiently than H_2. To see this, consider a primordial cloud at the "loitering" phase with $n_H \sim 10^4$ cm^{-3} and $T \sim 200$ K. At this point, radiative cooling by H_2 becomes inefficient, so the gas contracts only slowly, and fragmentation is suppressed at least until an accretion disk forms around the first protostar. This is why the characteristic mass of Population III.1 stars may be as high as $\sim 100\ M_\odot$ (see §5.1).

Now, let us imagine that the gas instead has a small fraction of metals; if these elements can efficiently cool the gas from this thermodynamic state, they will induce fragmentation to smaller mass scales. We use the common notation $[X/H] = \log_{10}(N_X/N_H) - \log_{10}(N_X/N_H)_\odot$ to describe the abundance of species X. Detailed calculations show that carbon and oxygen are the most important elements at the relevant temperature and density, at least for atomic cooling. Carbon is likely to be singly ionized C II, because the Universe is transparent to photons above the ionization potential of C I (11.26 eV), though it is sufficiently close to the Lyman-Werner bands that it may suffer some self-shielding by H_2 in very dense clumps (see also §4.6). Oxygen, in contrast, has an ionization potential very near that of H I (13.6 eV), and so it will remain neutral. Let us write $\Lambda_X(n, T)$ for the radiative cooling rate from species X, and Λ_{tot} for the total rate. For these two species at the relevant temperatures and densities, the cooling is dominated by fine-structure lines of O I (wavelength of 63.1 μm) and C II (157.7 μm).

AA x2

Fragmentation requires that the cooling time, $t_{cool} = 1.5 n k_B T / \Lambda_{tot}$, be smaller than the free-fall time in the gas, $t_{ff} \approx 1/\sqrt{G\rho}$. For a given species, this defines a *critical metallicity* $[X/O]_{crit}$ above which radiative cooling suffices to induce fragmentation. Detailed calculations of the fine-structure transitions in these elements yield $[O/H]_{crit} \approx -3.0$, and $[C/H]_{crit} \approx -3.5$, with a factor of ~ 2 uncertainty depending on the details of the thermodynamic state of the loitering phase.[17]

The preceding considerations include only gas-phase cooling; many of the ISM metals at low redshifts are contained in dust grains, which can also aid cooling due to both thermal emission and H_2 formation (which can occur very efficiently on the surface of dust grains, since hydrogen atoms became trapped in close proximity). Locally, dust formation is generally attributed to winds in asymptotic giant branch stars. At high redshifts, dust may be produced mainly in the metal-rich ejecta of supernovae themselves. The dust formed inside supernova ejecta is a very effective coolant, and some models show that the critical metallicity falls to $[Z/H]_{crit} \approx -6$ if such dust is produced efficiently.[18]

Regardless of its precise value, the small critical metallicity is easy to achieve. We have seen that a single pair-instability supernova remnant can easily fill an entire halo as well as some portion of the IGM. Typical explosions generate $M_{SN,C} \sim 10\ M_\odot$ or $\sim 30\ M_\odot$ of O. A single supernova therefore enriches its host to a carbon abundance $\sim 3 \times 10^{-3} (M_{SN,C}/10\ M_\odot)(10^6\ M_\odot/M_h)$ times the solar value (and to a comparable level for oxygen). Thus, provided only that mixing is efficient, a single supernova suffices to shift star formation in its host halo—and possibly its close neighbors—into the Population II channel.

The preceding arguments show that fragmentation can occur in low-metallicity environments, but they do not determine the actual spectrum of mass fragments. That distribution is highly uncertain, but it is still likely to be skewed to significantly higher masses than today. The arguments, as in §5.2.4, in which the CMB sets the temperature floor for the cooling gas, apply to these enriched clumps also, setting the characteristic mass scale to be a few tens of solar masses—still well into the high-mass regime.

Nevertheless, the transition to Population II was a crucial milestone in the history of the Universe. The arguments in this section suggest that if mixing was efficient, it took place very soon after the first star in each virialized halo exploded.

6.5.1 Blast waves in an Expanding Universe

A crucial point to understand about metal enrichment is that it *must* be highly inhomogeneous, because the metals are produced at discrete sites (star-forming halos) and must be advected with hydrodynamic flows, which typically move rather slowly by cosmological standards. Thus, the transition from Population III to Population II is likely to have had large spatial fluctuations; in principle, Population III star formation could have persisted to late times if the IGM enrichment timescales were very long and if new halos virialized and cooled in pristine gas. In this section, we consider how galactic winds (or other flows) can distribute this material around the Universe.

Although the simple models of §6.4.2 provide some intuition, they do not directly apply to cosmological blast waves, which propagate into an expanding medium whose density decreases with time. However, it is easy in this case to estimate the maximum distance to which the shock can reach: as in a uniform, static medium, the wave will sweep the matter before it into a thin shell. But in the cosmological setting, the shell will continue to expand in comoving coordinates only while its velocity *relative to the Hubble flow* is positive—after that, the shell will simply be dragged along with the expanding Universe. The drag from swept-up material will continue to decelerate the blast wave until its velocity matches the Hubble flow, which will occur at the asymptotic *proper* radius $R_{\rm E}$; if the expansion occurs quickly, the final kinetic energy is then $M_s[H(z_i)R_{\rm E}]^2/2$. (Here z_i is the redshift corresponding to the initial time of the explosion.)

Some of this energy comes from the explosion energy E, but in contrast with the static medium, the initial configuration also contains some kinetic energy from the Hubble flow. If we integrate outward to $R_{\rm E}$, this initial energy is $3M_s[H(z_i)R_{\rm E}]^2/10$. If we assume that the expansion is rapid compared with the Hubble time, the maximum size is therefore

$$R_{E,\rm com} \sim \left[\frac{E}{\rho_b(z_i)H^2(z_i)}\right]^{1/5}(1+z) \tag{6.39}$$

$$= K_{E,\cos}\left(\frac{GE}{H_0^4\Omega_b\Omega_m^2}\right)^{1/5}(1+z_i)^{-1/5}, \tag{6.40}$$

where $K_{E,\cos}\sim 10^{1/5}$ is a constant of the order of unity that depends on how much energy is transformed into thermal or kinetic energy. Note the similarity to the Sedov-Taylor–von Neumann scaling, with $t\sim 1/H(z)$.

In fact, for a perfectly adiabatic shock in a matter-dominated Universe with $\Omega_b \ll \Omega_m$, a self-similar blast wave that mirrors the Sedov-Taylor–von Neumann solution forms.[19] In this limit, the constant $K_{E,\cos} = (32\pi/3)^{1/5} K_{\rm STV}$. The blast wave also expands at a rate $R_{\rm com} \propto \tau^{2/5}$, where at high redshifts,

$$\tau(z) \approx \frac{2}{\sqrt{\Omega_m} H_0}[(1+z_i)^{1/2} - (1+z)^{1/2}] \quad (6.41)$$

Add period / PE

Once radiative cooling in the shell and eventually the bubble interior becomes important, the expansion slows down. We can estimate the final size of a bubble in which cooling is extremely efficient by repeating the preceding argument, but with momentum conservation as our guiding principle rather than energy conservation. Writing the total impulse as E/c, we obtain

$$R_{p,\rm com} \sim K_{\cos,p}\left[\frac{E/c}{\rho_b(z_i)H(z_i)}\right]^{1/4}(1+z) \quad (6.42)$$

$$= K_{p,\cos}\left(\frac{GE/c}{H_0^3\Omega_b\sqrt{\Omega_m}}\right)^{1/4}(1+z_i)^{-1/8}, \quad (6.43)$$

replace with subscript italics p, subscript cos / AA

where $K_{p,\cos}\sim 8^{1/4}$.

To put these estimates in the context of star-forming halos, we use the notation of §6.4.1 and write the energy released by a halo of mass M as $E = f_\star \omega_{\rm SN} M_g$ and the momentum input as $E/c = (\epsilon f_\star M_g c^2)/c$, where the gas mass is $M_g = (\Omega_b/\Omega_m)M_h$. Then,

$$R_{E,\rm com} \sim 1.2\left(\frac{\omega_{\rm SN}}{10^{49}\ {\rm erg}\ M_\odot^{-1}}\frac{f_\star}{0.1}\frac{M_h}{10^8\ M_\odot}\right)^{1/5}(1+z)^{-1/5}\ {\rm Mpc}, \quad (6.44)$$

$$R_{p,\rm com} \sim 0.2\left(\epsilon_3\frac{f_\star}{0.1}\frac{M_h}{10^8\ M_\odot}\right)^{1/4}(1+z)^{-1/8}\ {\rm Mpc}. \quad (6.45)$$

It follows that the maximal *volume* enriched by halos scales as $V_E \propto (f_\star M)^{3/5}$, or $V_p \propto (f_\star M)^{3/4}$. In either case, the scaling is sublinear, which shows that low-mass halos are much more efficient at enriching the IGM than massive ones.

In practice, these maximal radii can be substantial overestimates for two reasons. First, they neglect the gravitational attraction of the host halo. Second, the expansion must occur slowly: if the deceleration is modest, a shell with an initial speed v_i will take a time $t\sim R/v_i\sim(v_i/H)/v_i\sim 1/H$ to reach the radius at which that velocity matches the Hubble flow. Thus, using $H(z_i)$ in the estimates is not formally correct.

To follow the time evolution in detail one must track the energy reservoir driving the wind. Numerical calculations show that cosmological blast waves develop shells even more rapidly than their counterparts in static media. The

equation of motion for a shell is then

$$\ddot{R} = \frac{4\pi R^2}{M_s}(p - p_{\rm IGM}) - \frac{G}{R^2}[M(R) + M_s/2] + \Omega_\Lambda(z)H^2(z)R - \frac{\dot{M}_s}{M_s}(\dot{R} - HR), \tag{6.46}$$

where M_s is the shell mass, $\dot{M}_s = 4\pi R^2 \rho_b(\dot{R} - HR)$ is the rate at which mass is swept up, p is the pressure of the bubble interior, $p_{\rm IGM}$ is the ambient pressure of the IGM, and $M(R)$ is the mass enclosed within the wind (including both dark matter and any baryonic remnants). The first term is the pressure force from the hot interior, the second involves the gravitational deceleration due to the interior mass (and the shell itself), the third is the acceleration due to the cosmological constant (which can be ignored at high redshifts), and the last is the drag force from swept-up material. Equation (6.46) must be supplemented with an equation for the energy of the bubble interior,

$$\dot{p} = \frac{L}{2\pi R^3} - 5p\frac{\dot{R}}{R}. \tag{6.47}$$

Here the last term is the pdV work from expanding the shell, while the first represents energy inputs or losses. These include the energy source powering the wind, and Compton cooling (which usually dominates at the low bubble densities once the winds propagate into the IGM) or any other radiative process.

Insert space before period / PE

There is one additional subtlety in the cosmological case: the shell treatment assumes that the ambient gas is accelerated to the shell velocity through inelastic collisions of its particles. In reality, L must also account for the energy dissipated in this process. If the shell cooling time is short, most of it will be lost (the static medium solutions described in §6.4.2 implicitly take this limit), but some energy may be transmitted to the bubble interior through turbulence if it is not lost in cooling. We let f_d be the fraction of this energy transmitted to the bubble interior; then L includes a term

$$L_d = f_d \dot{M}_s(\dot{R} - HR)^2/2. \tag{6.48}$$

Figure 6.9 shows solutions to these shell expansion equations for halos with $M_h = 2\times10^6\ M_\odot$ beginning at a variety of redshifts (for ease of comparison, the IGM is assumed neutral and cold in all the curves). In all cases we assume an instantaneous burst of star formation with $f_\star = 1$ and take $f_d = 1$ and $f_d = 0$ for the upper and lower curves, respectively. Expansion is truncated when the bubble interior cools to low temperatures (i.e., at the end of the pressure-driven snowplow phase). Note the fairly long times it takes the bubbles to reach these limiting sizes.

Figure 6.10 shows the proper radii of wind bubbles surrounding halos (as a function of mass) at several different redshifts. The points use a Monte Carlo model of each halo's merger history (generated following the excursion set formalism of §3.4.2) to trace the supernova histories of each halo; the variations in the wind radius at a given mass therefore reflect variations in the star formation histories of the galaxies. The model adopts $f_\star = 0.1$ and $f_d = 0$, and also assumes that 75% of the supernova energy is radiated away without contributing to the wind. The solid lines show the virial radii of the host halos.

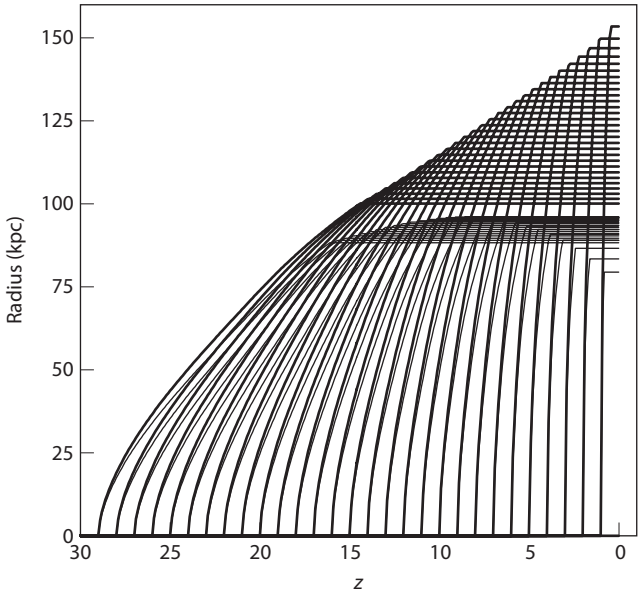


Figure 6.9 Shell sizes as a function of redshift, in comoving distance, from galaxies with $M_h = 2 \times 10^6\ M_\odot$ forming at redshifts from $z = 29$ to $z = 1$. The model assumes a star formation efficiency near unity. The upper set of curves take $f_d = 1$, while the lower set take $f_d = 0$. The expansion is assumed to stop when the bubble interior falls below $T = 15{,}000$ K. Tegmark, M., Silk, J., & Evrard, A., *Astrophys. J.* **417**, 54 (1993). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~ by / AAS. /

Clearly, the final wind sizes are only a weak function of halo mass, with $R \propto M_h^{0.2-0.25}$, as expected from the analytic scaling. The magnitudes are considerably below the maximal estimate from equation (6.44), because much of the energy is assumed to be radiated away, the Hubble flow energy is assumed unavailable, and the energy is not injected at a single instant. The results are, however, reasonably close to the momentum limit of equation (6.45).

The numerical results turn over at high masses, because the gravitational potential well of the host traps the wind. Typically, this occurs before the wind escapes far into the IGM, so there is a severe cutoff in the maximum size—recall that the gravitational binding energy scales as M_h^2, while the available energy scales only as M_h. This, together with the $V \propto E^{3/5} \propto (f_\star M_h)^{3/5}$ scaling of the enriched volume, means that the smallest halos are likely the most important ones for chemical enrichment, unless the star formation efficiency itself decreases strongly at low halo masses.

DES: 2 lines under head ok?

6.5.2 Metals in the Intergalactic Medium

Given the fate of a wind bubble around any individual source, it is straightforward to estimate the fraction of space filled by these bubbles. Defining $V(M_h, z)$

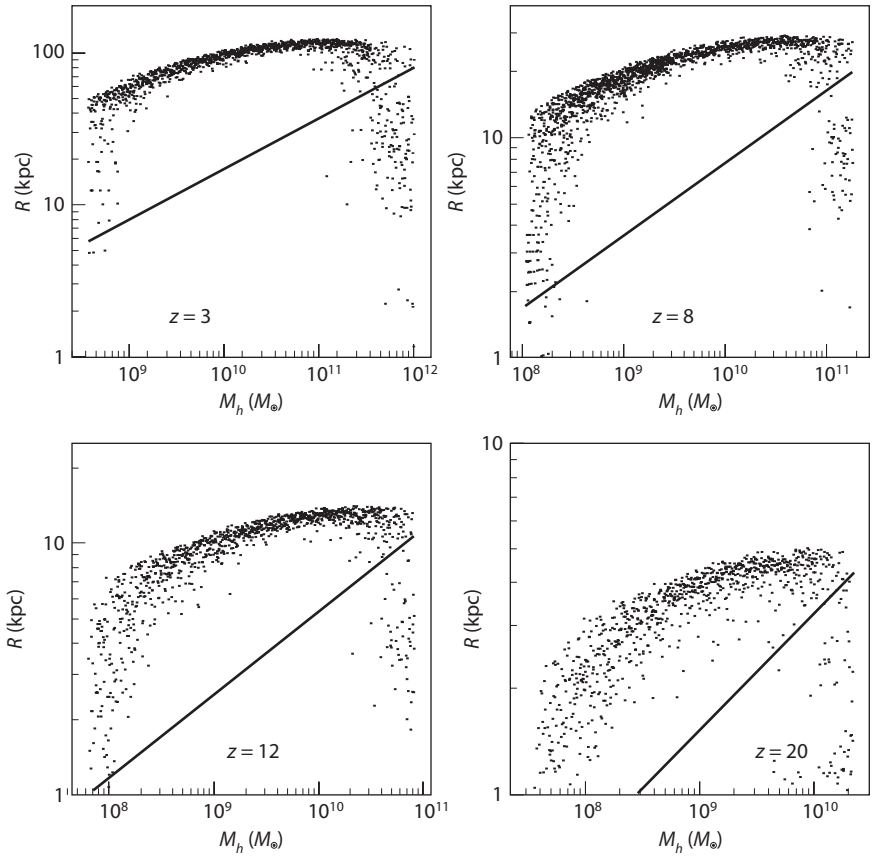


Figure 6.10 Proper radius of wind outflows at various redshifts, as a function of halo mass. Each point is a realization of a merger-driven star formation model (see text). The solid line shows the virial radius of the host halo. All the points assume that $f_\star = 0.1$, 75% of the supernova energy is radiated away without contributing to the wind, and $f_d = 0$. Note that the mass and radius scales change between panels. Furlanetto, S. R., & Loeb, A., *Astrophys. J.* **588**, 18 (2003). Reproduced with permission of the American Astronomical Society.

by / AAS. /

to be the volume filled by a bubble blown by a halo of mass M_h at redshift z,[viii] we integrate over the halo mass function:

$$Q'_e(z) = \int_{M_{\rm min}}^{\infty} dM\, n(M,z) V(M,z), \tag{6.49}$$

[viii] In reality, such a relationship will not be one-to-one, as halos (even of the same mass) form and grow with different merger and star formation histories. Figure 6.10 demonstrates this point explicitly.

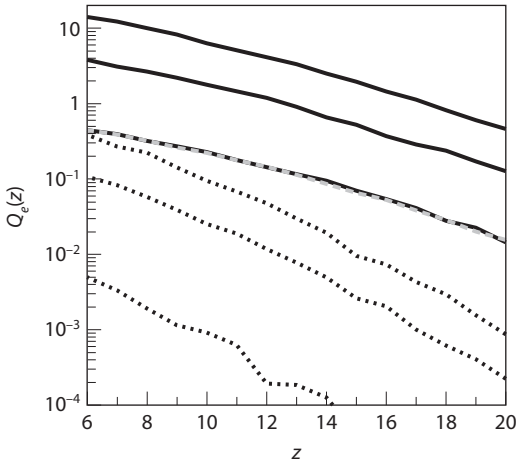


Figure 6.11 Filling factor of wind-enriched regions in different models of star formation and wind expansion, produced using the same Monte Carlo methods as in Figure 6.10. The solid and dotted sets of curves show models with and without efficient star formation in halos below the atomic cooling threshold (in all halos with $T_{\rm vir}>400$ K); otherwise, we allow star formation only in halos above the atomic cooling threshold ($T_{\rm vir}>10^4$ K). Within each set, the curves take $f_\star = 0.5$, 0.1, and 0.01, from top to bottom. For the H_2 cooling case, we assume that $\omega_{\rm SN} = 10^{50}$ erg $M_\odot^{-1}$ to reflect the powerful supernovae of very massive Population III.1 stars. The dashed curve (nearly coincident with the lowest solid curve) shows the H_2 case with $f_\star = 0.1$ and $\omega_{\rm SN} = 10^{49}$ erg $M_\odot^{-1}$. Furlanetto, S. R., & Loeb, A., *Astrophys. J.* **588**, 18 (2003). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

where the integration extends over all star-forming halos. The resulting Q'_e is the total volume filled by all the bubbles, not accounting for overlap. If the bubbles were randomly distributed, and if overlapping winds did not aid one another's expansion, the true filling fraction of wind material would be $Q_e = (1 - e^{-Q'_e})$.

This simple estimate has an important shortcoming: it ignores the clustering of these galaxies. In reality, high-redshift galaxies form close to one another along intersections of sheets and filaments in the cosmic web. Their wind bubbles therefore tend to overlap rather than fill new space. Because $V \propto E^{3/5}$, multiple sources contributing to a single bubble are *less* efficient than individual sources generating their own bubbles, so clustering will tend to decrease the filling fraction of the enriched material.

Figure 6.11 shows some example enrichment histories produced following the same methods as in Figure 6.10 but ignoring clustering. We consider two different sets of models: a maximal case that allows star formation in all halos with $T_{\rm vir}>400$ K (solid curves) and a more conservative one in which star formation occurs only in halos able to cool through atomic hydrogen transitions, with $T_{\rm vir}>10^4$ K. Within each set, the three curves assume $f_\star = 0.5$, 0.1, and

0.01, from top to bottom. In the small-halo case—which necessarily has to assume that H_2 cooling is extremely efficient—we also set $\omega_{\rm SN} = 10^{50}$ erg $M_\odot^{-1}$, which is 10 times larger than the nominal value for normal star formation, to reflect the large kinetic energy output expected from pair-instability supernovae (see §5.5). We also take $f_d = 0$ in this case; setting $f_d = 1$ modestly increases the filling factor. All these curves were generated through a Monte Carlo model that accounts for the star formation histories of individual halos, and so they are not necessarily smooth.

Figure 6.11 shows that for a large fraction of space to be filled with heavy elements by $z\sim6$, most of those metals must come from the shallow potential wells of very small halos, which also must produce stars very efficiently. Indeed, if supernova and photoionization feedback is as efficient as our earlier estimates suggest, it seems implausible to expect such halos to be able to convert even 10% of their baryons into stars. Thus, metal enrichment in these early phases seems likely to be very patchy, with important consequences for structure formation (see §6.6).

In the galaxies that were likely responsible for most of the metal enrichment, both supernova winds and radiation pressure from hot stars contributed to powering the outflows. The former ultimately provided more energy for the outflow, but much of that energy may have been lost as the supernova blast waves propagated through the dense ISM of the galaxies. The momentum inputs from the two channels are comparable for typical IMFs, so even if supernova remnant cooling is efficient, winds from starbursts should be able to enrich a small percentage of the IGM at high redshifts.

Unfortunately, numerical cosmological simulations currently lack the dynamic range to model self-consistently the launch of these winds and their propagation through the IGM (partly because the shells cannot be resolved, but, more important, because our physical understanding of winds is still fairly crude), although simulations of individual galaxies are beginning to examine outflow dynamics in detail. In large-scale structure simulations, winds are launched by hand with a parameterized model; their propagation through the IGM is then tracked in the momentum-dominated limit. Such numerical simulations also show that plausible models for winds from halos above the 10^4 K cooling threshold can enrich only a small percentage of the IGM. As the winds continue to expand at later times, this fraction increases, but many models predict that much of the IGM remains pristine even to late times.

The mean metallicity of these enriched regions follows easily from the preceding models with only one additional parameter: the fraction $f_{\rm met}$ of the galaxy's metals that are ejected in the wind. This variable is usually parameterized with the *mass-loading factor* η, which describes how much material escapes the galaxy in units of the star formation rate, $\eta = \dot{M}_w/\dot{M}_\star$. According to the model described in §6.4.1, $\eta\sim\sigma_0/\sigma$ (see equation 6.33), assuming that the momentum input rate simply scales with the star formation rate and that the final velocity of the winds is just a few times the escape velocity of the halo. Observations of low-redshift starbursts are consistent with this simple relation if $\sigma_0\sim300$ km s^{-1}, though we note that the proportionality constant depends on

the IMF and may get *larger* if the IMF is top-heavy at high redshifts. However, this, parameter provides yet another reason why small galaxies more efficiently enrich the IGM with metals, as $\eta \propto \sigma^{-1} \propto M_h^{-1/3}$.

Assuming that the metals are perfectly mixed inside the galaxy this implies that a fraction $\eta f_\star$ of the metals produced in each galaxy are ejected into the IGM. This material is then diluted by a factor $\sim Q_e f_{\rm coll}$ as it spreads into the IGM; thus, the mean IGM metallicity will be

$$Z_{\rm IGM} \sim 10^{-3} \langle \eta \rangle \left(\frac{f_\star}{0.1} \frac{f_{\rm coll}}{0.01} \right) Z_{\rm gal}, \tag{6.50}$$

where $Z_{\rm gal}$ is the mean metallicity of material inside galaxies, and $\langle \eta \rangle$ is averaged over the entire galaxy population. The mean metallicity of *enriched* regions will be larger by $\sim Q_e^{-1}$. Because Q_e should increase with $f_{\rm coll}$, this result shows that the metallicity of enriched regions is likely to be above the critical threshold for the transition to Population II star formation in most plausible scenarios.

An alternative empirical estimate of the IGM metallicity follows by observing the total density of stars (which, assuming an IMF, translates into a total metal yield). Type II supernovae from high-mass stars forming in a typical Salpeter IMF process $\approx$2.4% of the stellar mass into metals. If we use the observed stellar mass estimates at $z \sim 2$, this result implies that the IGM should have $Z \sim (1/30)\ Z_\odot$ at that redshift.

However, only $\sim$10% of these metals predicted by measuring the stellar mass of the Universe have actually been observed; the remainder may be buried inside additional galaxy populations or in diffuse IGM systems where the metal-line optical depth is too small to resolve. In the latter case, the enrichment may indeed be widespread, at least by $z \sim 2$–3.

For a similar constraint at higher redshifts, we can calibrate the stellar mass to the number of ionizing photons produced per baryon, which will let us gauge the overall level of enrichment near the time of reionization (see chapter 9). We let $Q_{\rm H\ II}$ be the number of ionizing photons reaching the IGM per hydrogen atom, $Q_{\rm H\ II} \approx N_\gamma f_{\rm esc} f_\star f_{\rm coll}$, where N_γ is the number of ionizing photons produced per baryon in stars ($\sim$4,000 for a Salpeter IMF), and $f_{\rm esc}$ is the fraction of these photons that escape their host galaxy into the IGM; we discuss these parameters in more detail in chapter 9. Meanwhile, the mean metallicity implied by these stars is $Z/Z_\odot \sim 1.3 f_\star f_{\rm coll}$, where the factor 1.3 is the conversion from the 2.4% metal yield to solar metallicity (which has 1.89% of the mass in metals). Thus,

$$Z \sim 3 \times 10^{-3} Q_{\rm H\ II} \left(\frac{400}{N_\gamma f_{\rm esc}} \right) Z_\odot. \tag{6.51}$$

Again, the mean metallicty in enriched regions will be a factor of Q_e^{-1} larger.

Our primary tools for constraining these winds are metal-line systems in the Lyman-α forest (see §4.6). Metals seem *nearly* ubiquitous in the high–column density systems that may be associated with virialized objects, which implies

AU: Add callout for Table 6.2 - see next page

Table 6.2 Summary of Feedback Processes Affecting the First Stars and Galaxies.

Type	Mechanism	Effect
Radiative	Lyman-α photons	Pressure-driven winds
	Lyman-Werner photons	H_2 photodissociation/ population III.2 stars
	Ionizing photons	Photoevaporation of halo gas, gas entropy increase (see §9.9)
	X-ray photons	Free-electron formation, gas entropy increase (see §9.9)
Mechanical	Supernovae	Halo disruption
	Superwinds	Gas suppression, metal pollution
	AGN winds	Gas suppression (see §7.5)
Chemical	Enrichment	Population II stars

that such halos are highly enriched.[ix] This is not surprising, since the first stars in any halo are themselves likely to enrich the hosts' material to substantial levels. More interesting is the wide scatter in the metallicity of lower-density regions. The estimate in equation (6.50) is reasonably close to the observed metallicities of these systems ($Z \sim 10^{-3} Z_\odot$), so careful studies of IGM metal lines over time may shed light on winds and other outflows. In particular, even at $z > 6$, these enriched regions will produce measurable absorption in quasar or GRB spectra, although identifying each line's origin may be difficult (see §4.7).

As we will see in §8, these winds likely also play a crucial role in regulating star formation within galaxies, and their parameters can therefore be estimated not only through IGM metallicity measurements but also by comparison with galaxy luminosity functions, metallicities, and other properties. This role provides another observational handle on winds and, indirectly, chemical enrichment processes.

6.6 The First Galaxies

In chapter 5 we discussed the physics of primordial star formation. Although there are many unanswered questions, the problem of first *star* formation is a tractable one: the initial conditions are well-posed and the physics (dark matter and baryonic collapse, chemistry of the primordial gas, accretion disk formation, and radiative feedback) is straightforward enough that one can at least imagine solving the problem in full.

[ix] For a counter example, see Fumagilli, M., O'Meara, J. M. & Prochaska, J. X., *Science* **334**, 1245 (2011).

, as summarized in Table 6.2 / AA

In this chapter we have examined the myriad feedback mechanisms generated by these stars and their descendants. As soon as the first stars formed, these processes complicated matters immensely, and it is extremely difficult to imagine building a picture of the subsequent generations of star formation from first principles—there are simply too many uncertain parameters driving each one. Nevertheless, the underlying physics of each process is relatively straightforward, and from detailed studies of each individual process we can build some intuition for how the interplay may have proceeded.

Such "global" formulations are coming into focus for the transformation of the first stars to the first *galaxies*. We define a "galaxy" as a gravitationally-bound system of stars embedded in a dark matter halo and exhibiting *sustained* star formation (even if at a low level) over cosmological time periods (i.e., a substantial fraction of the Hubble time), either in the past or ongoing at present. This definition requires (i) a virialized dark matter halo able to accrete baryons (hence $M > M_{\rm fil}$); (ii) efficient cooling in the baryons (above a critical virial temperature $T_{\rm min}$ that depends on the chemistry of the constituent gas); (iii) sufficient mass to be stable against feedback from its own stars; and (iv) sufficient mass to be stable against feedback from neighboring halos.

Here we describe a plausible scenario for how such objects can appear at high redshifts. It should be obvious, however, that though this represents a best guess given present theoretical investigations, the lack of observational constraints likely means that it is at best partially correct. Nevertheless, it provides a coherent synthesis of the concepts we have discussed and is a useful baseline paradigm for future work. Figure 6.12 illustrates the following evolutionary stages graphically and identifies some of their key points:

1. The first stars form inside halos cooled by molecular hydrogen, with characteristic masses determined by the chemistry of H_2 cooling (see §5.1.2 and Fig. 5.5) such that $T_{\rm vir} > 1{,}000$ K. Massive Population III.1 stars form at the center of these halos after cooling to low temperatures. The key question is whether the gas cloud fragments before the material accretes onto the protostar. If not, the final mass is likely regulated by radiative feedback (with $M_\star > 100\ M_\odot$); otherwise, the first protostar's accretion disk is the most likely site for fragmentation, and the characteristic mass may be several times smaller.
2. The first star (or star cluster) exerts extremely strong feedback on its host halo's gas. The H II region created by a very massive Population III.1 star evaporates any diffuse gas in the central regions of the halo (§6.3.1), and the star's death as a supernova triggers a blast wave that quickly clears out the rest of the gas (provided, of course, the star does not collapse directly to a black hole without generating an explosion; §6.4.3), and it enriches the entire halo with heavy elements. Dense clumps well on their way to star formation may survive this feedback (and, in fact, the shock compression may even speed up their collapse), but nothing else will. The feedback will be less severe, but still substantial, if Population III.1 stars are less massive. Nevertheless, *Population III.1 star*

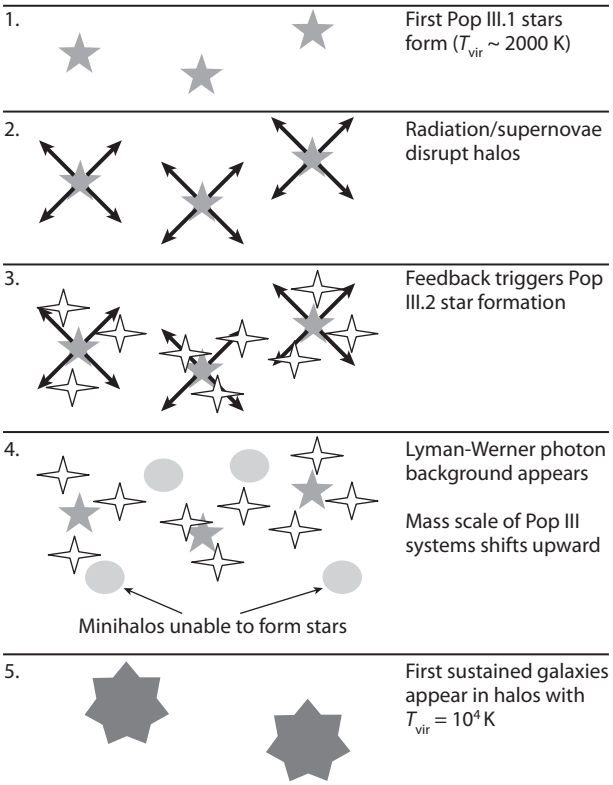


Figure 6.12 Stages in a plausible scenario for the birth of the first stars and galaxies (see text for details, and *Color Plate 10* for a color version of the figure). *(1)* The first Population III.1 stars form in small halos via H_2 cooling. *(2)* These stars empty their hosts of gas via photoevaporation and supernova blast waves. *(3)* This feedback triggers Population III.2 star formation in nearby minihalos. *(4)* The Lyman-Werner background from these stars suppresses star formation in small minihalos, gradually increasing the characteristic mass scale of star-forming objects. *(5)* The first self-sustaining galaxies eventually form in halos above the atomic cooling threshold, $T_{vir} \sim 10^4$ K.

formation in any individual halo may occur only in a single rapid burst.

3. These same feedback mechanisms also operate on somewhat larger scales, as the H II region and supernova blast wave are able to penetrate to approximately kiloparsec scales. Any nearby halos are therefore subject to the same effects: baryons in those that have not yet collapsed to high densities evaporate at high entropies, while star formation is likely accelerated in those already dense enough to self-shield from the ionizing radiation (§6.3.1). However, because these systems form their stars from ionized gas, the enhanced HD chemistry leads to more efficient cooling and hence (probably) a smaller characteristic mass of Population III.2 stars (see §5.3).

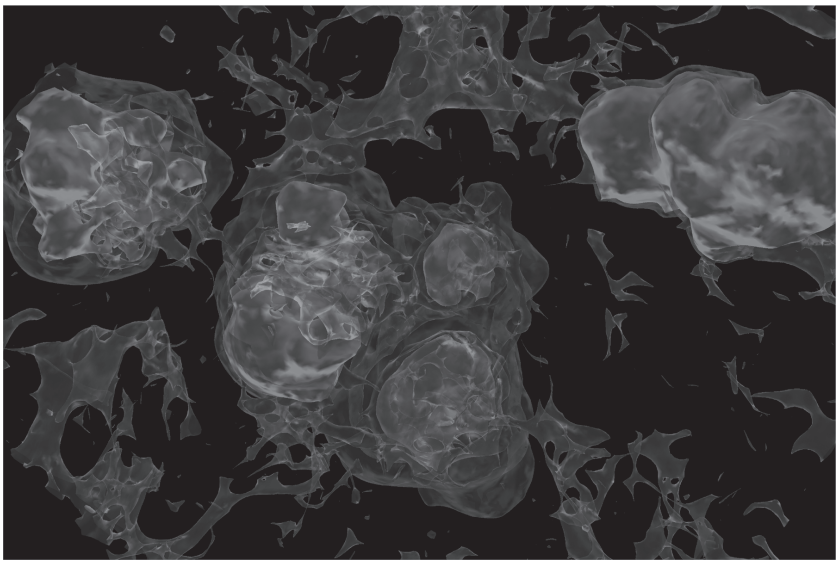


Figure 6.13 Results from a numerical simulation of the formation of a metal-free stars and their feedback on the surrounding environment (see *Color Plate 11* for a color version of the figure). Radiative feedback around the first star involves ionized bubbles (~~light~~ gray) and regions of high molecule abundance (~~medium~~ gray). The large residual free-electron fraction inside the relic ionized regions, left behind after the central star died, rapidly catalyzes the reformation of molecules and a new generation of lower-mass stars. ~~Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F., *Nature* **459**, 49 (2009). Copyright 2009 by Nature Publishing Group.~~

medium / light / AA

(Note that because supernova blast waves travel much slower than H II regions, it is very possible for this triggered star formation to be metal free.) Still, even with this positive feedback, the Population III.1 and Population III.2 stars in a given cluster of minihalos form close together temporally (as otherwise, the clumps would have photoevaporated), leading to "bursts" of Population III stars followed by long pauses as the halos reaccrete their gas. Figure 6.13 illustrates some of the complexity of this stage: note the several nearby stars that form and the complicated morphology of the molecular gas catalyzed by the presence of the H II regions.

4. Feedback also operates on larger scales. All Population III stars produce photons in the Lyman-Werner that photodissociate H_2. As more stars form, the Lyman-Werner background increases, gradually raising the critical virial temperature for cold gas formation inside minihalos (§6.1.5 and Figure 6.4). Because more massive halos are also more rare, this tends to self-regulate the global rate of star formation.
5. Eventually, the Lyman-Werner background becomes intense enough to choke off Population III star formation in pristine minihalos entirely. Then, star formation shifts to halos with $T_{\rm vir} > 10^4$ K, where H I is

Reprinted from *Nature*, 459, 7243, Bromm, V., N. Yoshida, L. Hernquist, & C.F. McKee, "The formation of the first stars and galaxies," Figure 4, Copyright 2007, with permission from Nature Publishing Group.

ionized by the virial shock, and atomic cooling is efficient (see Figure 5.1). Most likely these halos had progenitors that formed Population III stars, in which case they already preenriched with metals and begin to form Population II stars. It is possible, however, that some such halos form without stars inside their progenitors (perhaps because they form relatively late and so star formation is suppressed by the Lyman-Werner background). In that case, their pristine, initially ionized gas triggers Population III.2 star formation.

6. Systems with $T_{\rm vir} > 10^4$ K can also maintain reasonable (though still small) star formation rates without completely disrupting their gas supplies (see equation 6.26). *It is therefore this "second generation" of star-forming halos that hosts the first sustained galaxies.*
7. Nevertheless, feedback continues to be important in regulating galaxy formation at later times. Winds and outflows are likely crucial for regulating star formation inside galaxies (see §8), and photoheating from ionizing photons in the IGM gradually increases the Jeans mass and thus the minimum mass scale for galaxy formation (see §9.9, where we discuss this topic in detail.)

The transition to star formation in long-lived galaxies likely occurred long before the Universe was reionized. The intensity of the Lyman-Werner background can be estimated as

$$J_{\rm LW} \sim \frac{c n_{\rm LW}}{4\pi} \left(\frac{h\nu}{\Delta \nu_{\rm LW}} \right) \tag{6.52}$$

where $n_{\rm LW}$ is the number density of photons in the Lyman-Werner band and $\Delta \nu_{\rm LW}$ is the bandwidth in frequency space. We then write $n_{\rm LW} \sim f_{\rm LW/ion} Q_{\rm H\ II} / f_{\rm esc} n_{\rm H}$, where $f_{\rm LW/ion}$ is the number of Lyman-Werner photons produced per ionizing photon by stars (which is $\sim$0.1 for very massive Population III stars, or near unity for Population II stars), $Q_{\rm H\ II}$ is the number of ionizing photons that escape into the IGM per hydrogen atom, and $f_{\rm esc}$ is the fraction of all ionizing photons that manage to escape in this way. Then,

$$J_{\rm LW,21} \sim 100 Q_{\rm H\ II} \left(\frac{0.1}{f_{\rm esc}} \frac{f_{\rm LW/ion}}{0.1} \right) \left(\frac{1+z}{20} \right)^3 . \tag{6.53}$$

Lyman-Werner photons suppress H_2 cooling completely when $J_{\rm LW,21} > 1$ (see equation 6.4), which should occur long before enough ionizing photons are produced to reionize the IGM. Thus, it seems very likely that the primary sources responsible for reionizing the Universe were long-lived galaxies rather than the bursty minihalos in which the first stars themselves formed.

Although this is a very plausible picture consistent with detailed theoretical work, there are a number of points at which seemingly minor differences may dramatically alter the results. We list several here to give a flavor for the uncertainties:

- If fragmentation is efficient in accretion disks composed of primordial stars, the first halos will form clusters of moderately sized stars rather

than single very massive stars. The resulting feedback will be less efficient, potentially allowing gas to remain in halos somewhat below the usual $T_{\rm vir}\sim10^4$ K atomic cooling threshold. The mass scale of the first galaxies would shift downward.

- If Population III stars form in the mass ranges 40–100 $M_\odot$ or 140–260 $M_\odot$, they will die by imploding rapidly to black holes without explosions. This would allow their halos to retain more of their gas, with only the photoevaporation feedback to contend with, and allow sustained star formation to continue in low-mass halos. Moreover, they would not enrich their environments (except perhaps weakly through stellar winds), allowing Population III.1 and III.2 to persist for longer timescales—possibly even to the atomic cooling threshold.
- If the shells that form at the edges of supernova blast waves are gravitationally unstable, they can fragment and form stars as well. If the fragmentation scale is small, these could even be long-lived stars that exert relatively small feedback.
- If black holes form abundantly and accrete gas efficiently from binary star companions or the ISM (see §7), then their X-ray background increases the free-electron fraction inside halos, promoting H_2 formation and possibly counteracting photodissociation from Lyman-Werner photons. This would allow much more rapid primordial star formation in low-mass halos.
- The consequences of enrichment inside minihalos has been largely unexplored, because the gas is expected to be expelled. But if some is retained, the metals allow rapid cooling and hence more efficient star formation than H_2. This, too, could lead to smaller galaxies.

Obviously, there is a great deal of uncertainty in how the first stars will grow into the first galaxies—most likely, observations will be necessary to settle the question. However, in closing we stress that most of the underlying physics is well understood in isolation and has many applications to other areas of astrophysics. It is the complex interplay of the processes we have described here that makes the problem challenging and exciting to explore observationally.

Chapter Seven

Supermassive Black Holes

7.1 Quasars and Black Holes: An Overview

A black hole is the end product of 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 of gravity, 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 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. More than 10 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 supermassive black holes that sit at the center of galaxies, reaching masses of up to 10 billion Suns. The latter type are observed as *active galactic nuclei* (AGN). It is by studying these accreting black holes that all our observational knowledge of black holes has been obtained.

A quasar—the most powerful type of AGN—is a point-like ("quasi-stellar") bright source at the center of a galaxy. Many lines of evidence indicate that a quasar involves a supermassive black hole that 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 toward the black hole and dissipates its rotational energy through viscosity. The gas is expected to drift inward in an accretion disk until it reaches the last possible stable orbit according to general relativity. Interior to this point, 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 is described in detail in §7.2, the fraction of the rest mass of the gas that gets radiated away just outside this orbit is high, ranging between 5.7% for a nonspinning black hole to 42.3% for a maximally spinning black hole (see Figure 7.5). This "radiative efficiency" is far greater

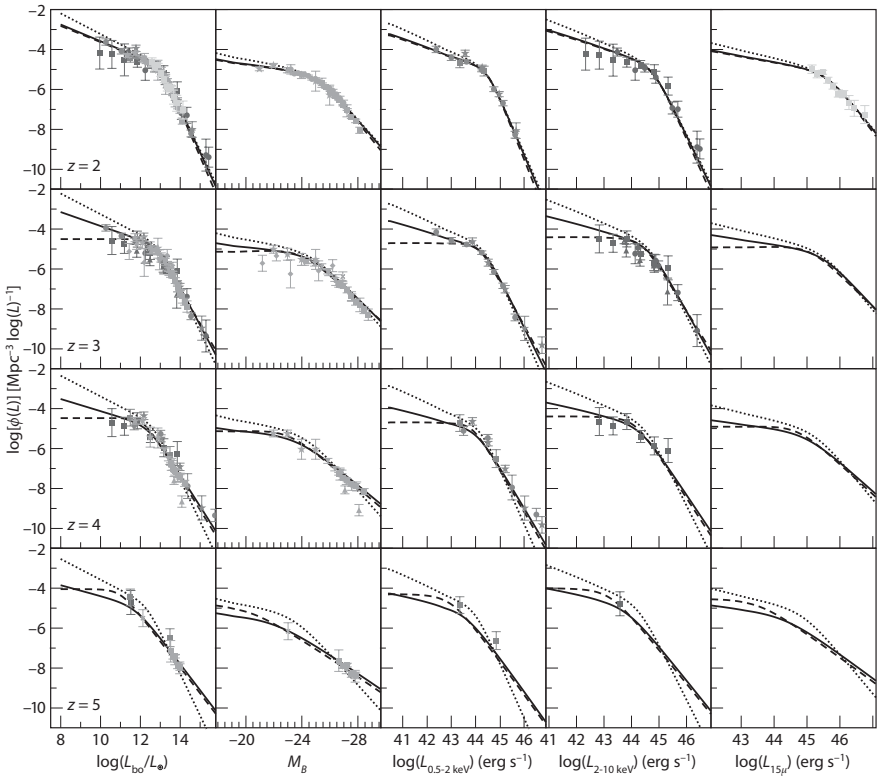


Figure 7.1 Redshift evolution of the luminosity function of quasars at different observed wavelengths: B-band (center-left panels), soft X-rays (0.5–2 keV) (center), hard X-rays (2–10 keV) (center-right, ~~red~~), and mid-IR (15 μm) (right, ~~cyan~~). The left panels show the distribution of bolometric luminosities (integrated over all wavelengths). Lines show the best-fit evolving double power-law model to data points at all redshifts (solid), the best-fit model at the given redshift (dashed), and the best-fit model that allows only the break luminosity to evolve (dotted). Hopkins, P. F., Richards, G. T., & Hernquist, L., *Astrophys. J.* **654**, 731 (2007). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

Delete (including semicolons) / AA

by / AAS. /

than the mass–energy conversion efficiency provided by nuclear fusion in stars, which is $< 0.7\%$.

Fortunately, quasars are very easy to see when the accretion occurs through a thin disk, and we have a great deal of demographic information on their properties out to very high redshifts. Quasar activity is observed in a small fraction of all galaxies at any cosmic epoch. Figure 7.1 shows the evolution of the luminosity function of quasars at different observed wavelengths in the redshift interval $z = 2$–5. Mammoth black holes weighing more than a billion solar masses have been discovered at redshifts as high as $z \sim 6$–7, less than a billion years after the Big Bang. The highest redshift quasar known (as of winter 2012) is ULAS J1120+0641 at $z = 7.085$ (only 0.77 Gyr after the Big Bang), with a bolometric luminosity of $6.3 \times 10^{13}\ L_\odot$ and an estimated black hole mass of $2 \times 10^9\ M_\odot$;

Insert comma / AA

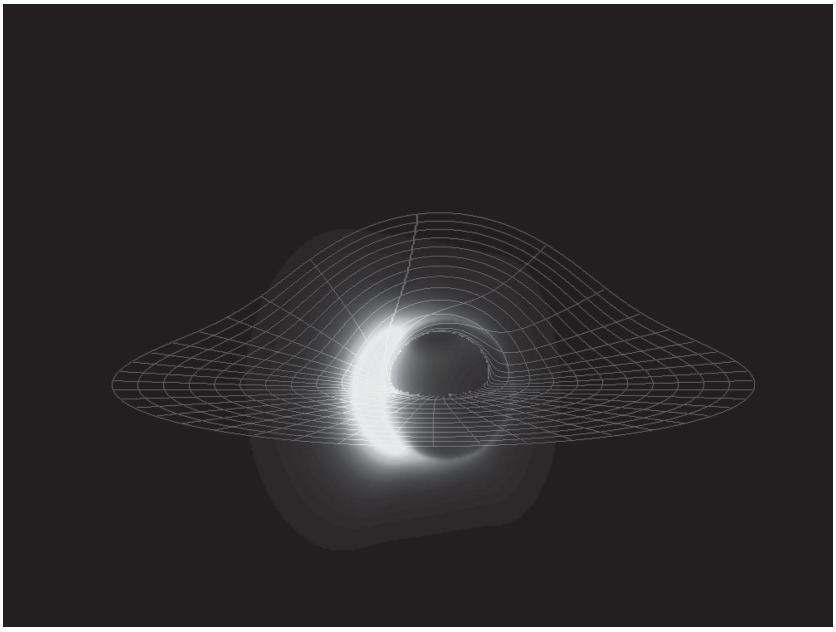


Figure 7.2 Simulated image of an accretion flow around a black hole spinning at half its maximum rate, from a viewing angle of 10° relative to the rotation axis (see *Color Plate 12* for a color version of this figure). The coordinate grid in the equatorial plane of the spiraling flow shows how strong lensing around the black hole bends the back of the apparent disk up. The left side of the image is brighter owing to its rotational motion toward 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, interferometers operating at millimeter wavelengths across the earth are necessary. Broderick, A., & Loeb, A. *Journal of Physics Conf. Ser.* **54**, 448 (2006); *Astrophys. J.* **697**, 1164 (2009). Reproduced with permission of the American Astronomical Society.

we show a partial spectrum of this quasar in Figure 11.8, where we discuss this object in the context of the reionization of the Universe.

There is clear and direct evidence for supermassive black holes, even beyond the AGN population. In our own Milky Way galaxy, stars are observed to zoom around the Galactic center at speeds of up to 10,000 km s^{-1}, owing to the strong gravitational acceleration near the central black hole (with a mass $\sim 4 \times 10^6\ M_\odot$).[1] But closer-in observations are forthcoming. Existing technology should soon be able to image the silhouette of the supermassive black holes in the Milky Way and M87 galaxies directly (see Figure 7.2).

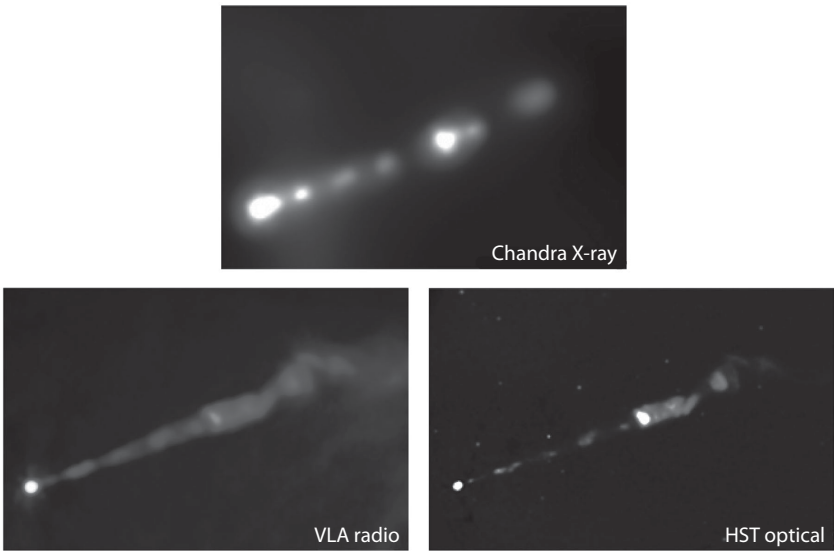


Figure 7.3 Multiwavelength images of the highly collimated jet emanating from the supermassive black hole at the center of the giant elliptical galaxy M87 (see *Color Plate 13* for a color version of this figure). The X-ray image (*top*) was obtained with the Chandra X-ray observatory, the radio image (*bottom left*) was obtained with the Very Large Array (VLA), and the optical image (*bottom right*) was obtained with the Hubble Space Telescope (HST). CXO/NASA.

Courtesy of /

Nevertheless, many questions remain about black holes. If the accreting material is organized into a thin disk, where the gas can efficiently radiate its released binding energy, then its theoretical modeling 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. For example, 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 crude understanding of black hole jets (see Figure 7.3) and other feedback mechanisms. These extraordinary exhibitions of the power of black holes are moving at nearly the speed of light and involve narrowly collimated outflows whose base has a size comparable to the solar system, while their front reaches scales comparable to the distance between galaxies.[2] Unresolved issues are as basic as what jets are made of (whether electrons and protons, electrons and positrons, or primarily electromagnetic fields) and how they are accelerated in the first place. The answers to these questions depend on critically on the role of the black hole spin in the jet-launching process. The

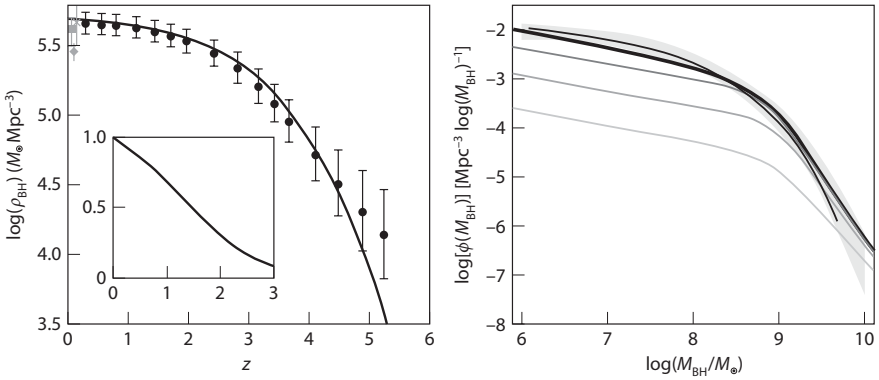


Figure 7.4 *Left panel:* The comoving black hole mass density of quasars from the data (circles) and the best-fit luminosity function (solid line). The inset shows the fraction of the mass density relative to today's value, $\rho_{\rm BH}(z)/\rho_{\rm BH}(0)$, on a linear scale. *Right:* The black hole mass function at $z = 0$ (thick black line) and $z = 1,\ 2,\ 3$ (from top to bottom). The shaded region shows the 1σ observational uncertainty. Hopkins, P. F., Richards, G. T., & Hernquist, L., *Astrophys. J.* **654**, 731 (2007). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

bright radiation and powerful jets of AGN can have dramatic feedback effects on the black hole's host galaxy.

If massive black holes grew 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, which implies that quasars are rare simply because their activity is short-lived. The inferred growth in the comoving mass function of black holes along with its integral over all black hole masses (i.e., the comoving mass density) are shown in Figure 7.4. Moreover, there appears to be a tight correlation between black hole mass and the gravitational potential-well depth of their host spheroids of stars (as measured by the velocity dispersion of these stars), as described in detail in §7.5.1.

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 prevents additional gas from feeding them further, or else the feedback from star formation in the vicinity of the black holes affects or controls their self-regulation. The black hole is like a baby who gets more energetic as he eats more at the dinner table, until his hyperactivity becomes 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.[3] Evidently, the growth of supermassive black holes is intimately linked to the hierarchical growth of their host galaxies. 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.

This state of affairs can easily be understood from the fact that the binding energy per unit mass in a typical galaxy correspond to velocities v of hundreds of kilometers per second 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.

Why did the collapsed matter in the Universe end up making galaxies and not black holes? One would have naively expected spherical collapse to end with the formation of a point mass at its center. But, as it turns out, tides from neighboring objects torque the infalling material and induce asphericity 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. In the previous two chapters we discussed how this material forms stars and, eventually, galaxies.

But star formation does not imply that no gas accumulates at the center. In fact, galactic spheroids are observed generically to harbor a central black hole, whose formation is most likely linked to the small mass fraction of galactic gas ($<0.1\%$) with unusually small 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.

In this chapter, we study formation mechanisms for supermassive black holes, their observable characteristics, and their interactions with their host galaxies and the wider Universe. We begin with a short introduction to the properties of black holes in general relativity.

7.2 Basic Principles of Astrophysical Black Holes

In Newtonian gravity, the gravitational field at any radius outside a spherical mass distribution depends only on the mass interior to that radius. This result is also true in Einstein's general relativity, where Birkhoff's theorem (see §1.2.2) states that the only vacuum, spherically symmetric gravitational field is that described by the static *Schwarzschild metric*,

$$ds^2 = -\left(1 - \frac{r_{\rm Sch}}{r}\right) c^2 dt^2 + \left(1 - \frac{r_{\rm Sch}}{r}\right)^{-1} dr^2 + r^2 d\Omega, \qquad (7.1)$$

where $d\Omega = (d\theta^2 + \sin^2\theta\, d\phi^2)$. The *Schwarzschild radius* is related to the mass M of the central (nonspinning) black hole,

$$r_{\rm Sch} = \frac{2GM}{c^2} = 2.95 \times 10^5 \text{ cm} \left(\frac{M}{1\, M_\odot}\right). \qquad (7.2)$$

The black hole horizon, $r_{\rm Hor}$ ($= r_{\rm Sch}$ here), is a spherical boundary from which 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 spatial region into which particles may fall but never come out breaks the 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 be characterized only by its spin J and electric charge Q (similarly to an elementary particle). In astrophysical circumstances, any initial charge of the black hole would quickly be 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 comparable to the gravitational force on a proton, $eQ \sim GMm_p$, which implies that $(Q^2/GM^2) \sim Gm_p^2/e^2 \sim 10^{-36}$ 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 Boyer-Lindquist coordinates:

$$ds^2 = -\left(1 - \frac{r_{\rm Sch} r}{\Sigma_k}\right) c^2\, dt^2 - \frac{2 j r_{\rm Sch} r \sin^2\theta}{\Sigma_k} c dt\, d\phi + \frac{\Sigma_k}{\Delta} dr^2 + \Sigma_k\, d\theta^2 + \left(r^2 + j^2 + \frac{r_{\rm Sch} j^2 r \sin^2\theta}{\Sigma_k}\right) \sin^2\theta\, d\phi^2, \tag{7.3}$$

where the black hole is rotating in the ϕ direction, $j = [J/Mc]$ is the normalized angular momentum per unit mass (in units of cm), $\Delta = r^2 - r r_{\rm Sch} + j^2$, and $\Sigma_k = r^2 + j^2 \cos^2\theta$. The dimensionless ratio $a = j/(GM/c^2)$ is bounded by unity, and $a = 1$ corresponds to a maximally rotating black hole. The horizon radius $r_{\rm Hor}$ is now located at the larger root of the equation $\Delta = 0$, namely, $r_+ = (1/2) r_{\rm Sch}[1 + (1 - a^2)^{1/2}]$. The Kerr metric converges to the Schwarzschild metric for $a = 0$. There is no Birkhoff's theorem for a rotating black hole.

Insert parentheses / PE

Test particle orbits around black holes can be simply described in terms of an effective potential. For photons around a Schwarzschild black hole, the potential is simply $V_{\rm ph} = (1 - r_{\rm Sch}/r)/r^2$, which leads to circular photon orbits at a radius $r_{\rm ph} = (3/2) r_{\rm Sch}$. For a spinning black hole,

Insert parentheses / PE

$$r_{\rm ph} = r_{\rm Sch}\left[1 + \cos\left(\frac{2}{3}\cos^{-1}[\pm a]\right)\right], \tag{7.4}$$

where the upper sign refers to orbits that rotate in the direction opposite that of 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 $r_{\rm ph} = (1/2) r_{\rm Sch}$ for a prograde orbit and $2 r_{\rm Sch}$ for a retrograde orbit.

Insert parentheses / PE

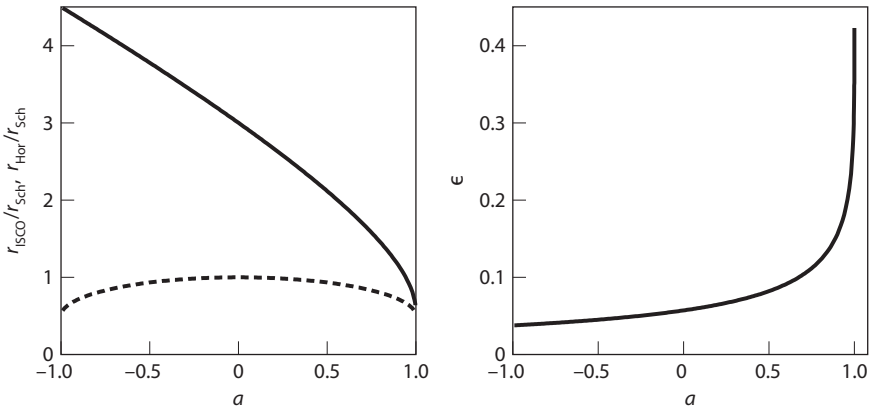


Figure 7.5 The left panel shows the radius of the black hole horizon $r_{\rm Hor}$ (dashed line) and the innermost stable circular orbit (ISCO) around it $r_{\rm ISCO}$ (solid line), in units of the Schwarzschild radius $r_{\rm Sch}$ (see equation 7.2), as functions of the black hole spin parameter a. The limiting value of $a = 1$ ($a = -1$) corresponds to a corotating (counterrotating) 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 in the right panel.

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 stable circular orbit (ISCO)* defines the inner edge of any disk of particles in circular motion (such as fluid elements in an accretion disk). At smaller radii, gravitationally bound particles plunge into the black hole on a dynamical time. This radius of the ISCO is given by[4]

$$r_{\rm ISCO} = \frac{1}{2} r_{\rm Sch} \left\{ 3 + Z_2 \pm [(3 - Z_1)(3 + Z_1 + 2Z_2)]^{1/2} \right\}, \tag{7.5}$$

where $Z_1 = 1+(1-a^2)^{1/3}[(1+a)^{1/3}+(1-a)^{1/3}]$, and $Z_2 = (3a^2+Z_1^2)^{1/2}$. Figure 7.5 shows the radius of the ISCO as a function of spin. The binding energy of particles at the ISCO defines their maximum radiative efficiency, because they spend a short time on their plunging orbit interior to the ISCO. This efficiency is given by

$$\epsilon = 1 - \frac{r^2 - r_{\rm Sch} r \mp j\sqrt{\frac{1}{2} r_{\rm Sch} r}}{r(r^2 - \frac{3}{2} r_{\rm Sch} r \mp 2j\sqrt{\frac{1}{2} r_{\rm Sch} r})^{1/2}}. \tag{7.6}$$

The efficiency ranges between $\epsilon = (1 - \sqrt{8/9}) = 5.72\%$ for $a = 0$, to $(1 - \sqrt{1/3}) = 42.3\%$ for a prograde (corotating) orbit with $a = 1$, and $(1-\sqrt{25/27}) = 3.77\%$ for a retrograde orbit.

7.3 Accretion of Gas onto Black Holes

7.3.1 Bondi Accretion

Consider a black hole embedded in a hydrogen plasma of uniform density $\rho_0 = m_p n_{\rm H}$ 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 can drive accretion of gas particles that are gravitationally bound to it, namely, interior to the radius of influence, $r_{\rm inf} \sim GM/c_s^2$. 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,

$$\dot{M} \approx \pi r_{\rm inf}^2 \rho_0 c_s = 15 \left(\frac{M}{10^8\ M_\odot}\right)^2 \left(\frac{n_{\rm H}}{1\ {\rm cm}^{-3}}\right) \left(\frac{T_0}{10^4\ {\rm K}}\right)^{-3/2} M_\odot\ {\rm yr}^{-1}. \quad (7.7)$$

In a steady state this supply rate equals the mass accretion rate into the black hole (see also §5.2.1).

The explicit steady-state solution to the conservation equations of the gas (mass, momentum, and energy), which is self-similar, was first derived by Bondi in 1952. This exact solution introduces a correction factor of the order of unity to equation (7.7).[5] Well inside the sonic radius (i.e., the point at which the infall and sound speeds are equal), the velocity is close to freefall $u \sim (2GM/r)^{1/2}$, 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 outward is much longer than the free-fall time. If the inflowing gas contains near-equipartition magnetic fields, then cooling through synchrotron emission typically predominates 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 is then $\sim GM/V^2$, which suggests that the sound speed c_s should be crudely replaced with $\sim (c_s^2 + V^2)^{1/2}$ in equation (7.7). A similar suppression factor applies for the accretion of baryons onto dark matter halos when the baryons have a net bulk velocity relative to the dark matter (see §2.1.2 and 3.2.2).

7.3.2 Thin-Disk Accretion

If the inflow is endowed with rotation, the gas will reach a centrifugal barrier from which it can accrete farther inward only after its angular momentum is transported away. This limitation follows from the steeper radial scaling of the centrifugal acceleration ($\propto r^{-3}$ if angular momentum is conserved; see equation 5.33) compared with the gravitational acceleration ($\propto r^{-2}$). Near the centrifugal barrier, where the gas is held against gravity by rotation, an accretion disk will form around the black hole, centered on the plane perpendicular to the rotation axis. The accretion time will then be dictated by the rate at which angular momentum is transported through viscous stress, and it could

be significantly longer than the free-fall time for a nonrotating flow (as in the solution described by the Bondi accretion model). As the gas settles to a disk, the dissipation of its kinetic energy into heat will make the disk thick and hot, with a proton temperature close to the gravitational potential energy per proton, $\sim 10^{12}\ \mathrm{K}(r/r_{\mathrm{Sch}})^{-1}$ in the absence of radiative processes. However, if the cooling time of the gas is shorter than the viscous time, then a thin disk will form. This latter regime is realized for the high gas inflow rate during the processes (such as galaxy mergers) that feed quasars. To better understand such objects, we start by exploring the structure of thin disks that characterize the high accretion rate of quasars.[6]

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 $v_\phi = r\Omega = (GM/r)^{1/2}$, where Ω is the angular velocity, and spirals slowly inward with radial velocity $v_r \ll v_\phi$ as viscous torques transport its angular momentum to the outer part of the disk. The associated viscous stress generates heat, which is radiated away locally from the the disk surface. We assume that the disk is fed steadily, and so it manifests a constant mass accretion rate at all radii. Mass conservation implies that

$$\dot{M} = 2\pi r \Sigma v_r = \text{constant}, \tag{7.8}$$

where $\Sigma(r)$ is the surface mass density of the disk.

In the limit of a 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 one another. The concept of shear viscosity can easily be 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 transported per unit time, $\sim nvm\Delta V$, is linear in the velocity gradient $\eta\, dV_y/dx$, with a viscosity coefficient $\eta \sim \rho v\lambda$ (in $\mathrm{g\,cm^{-1}\,s^{-1}}$), where $\rho = mn$ is the mass density of the gas. Since the excess kinetic energy density across a mean free path, $(1/2)\rho(\lambda\, dV_y/dx)^2$, is dissipated every collision time $\sim (\lambda/v)$, viscosity heats the gas at a rate per unit volume of $\mathcal{H}_{\mathrm{vis}} \sim [\eta(dV_y/dx)]^2/\eta$.

Insert parentheses / PE

Similar arguments show that within a Keplerian accretion disk, the flux of ϕ-momentum transported in the positive r-direction is given by the viscous stress $f_\phi = (3/2)\eta\Omega$. This dissipation is expected to be effective down to the ISCO, from which the gas plunges into the black hole over a free-fall time. We therefore set the inner boundary of the disk as r_{ISCO}, depicted in Figure 7.5. Conservation of angular momentum requires that the net rate of its change within a radius r be

Insert parentheses / PE

equal to the viscous torque, namely,

$$f_\phi \times (2\pi r \times 2h) \times r = \dot{M}\left[(GMr)^{1/2} - (GMr_{\rm ISCO})^{1/2}\right]. \tag{7.9}$$

The production rate of heat per unit volume by the viscous stress is given by $\mathcal{H}_{\rm vis} = f_\phi^2/\eta$. Substituting for f_ϕ in equation (7.9), we have

$$2h\mathcal{H}_{\rm vis} = \frac{3\dot{M}}{4\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.10}$$

This power provides a local radiative flux that leaves the system vertically from the top and bottom surfaces of the disk,

$$F = \frac{1}{2} \times 2h\mathcal{H}_{\rm vis} = \frac{3\dot{M}}{8\pi r^2}\frac{GM}{r}\left[1 - \left(\frac{r_{\rm ISCO}}{r}\right)^{1/2}\right]. \tag{7.11}$$

The total luminosity of the disk is given by

$$L = \int_{r_{\rm ISCO}}^{\infty} 2F \times 2\pi r\, dr = \frac{1}{2}\frac{GM\dot{M}}{r_{\rm ISCO}}, \tag{7.12}$$

where we have ignored general-relativistic corrections to the dynamics of the gas and the propagation of the radiation it emits.

In the absence of any vertical motion, momentum balance in the vertical z-direction yields

$$\frac{1}{\rho}\frac{dp}{dz} = -\frac{GM}{r^2}\frac{z}{r}, \tag{7.13}$$

where $z \ll r$ and p and ρ are the gas pressure and density, respectively. This equation fixes the 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 particle collisions, the particle-level viscosity is negligible in accretion disks. However, such disks are susceptible to the powerful *magnetorotational instability* (MRI) that amplifies magnetic turbulence over an orbital time. The origin of the instability can easily be understood by imagining two fluid elements that are threaded by a single magnetic field line and are slightly displaced from each other in the radial direction. The magnetic field acts as a spring owing to its tension. In a Keplerian disk the inner fluid element orbits more rapidly than the outer element, causing the spring to stretch. The inner fluid element is then forced by the spring to slow down, to reduce its angular momentum, and therefore to move to a lower orbit. The outer fluid element, meanwhile, is forced by the spring to speed up, to increase its angular momentum, and therefore to move to a higher orbit. The spring tension increases as the two fluid elements separate farther, and eventually the process runs away. The magneto-rotational instability[7] is likely to develop turbulent eddies in the disk that are much more effective at transporting its angular momentum than particle viscosity. In this case λ and v should be replaced by the typical size and velocity of an eddy. The largest value that these variables can attain are the scale height h and sound speed c_s in the disk, which implies

Delete hyphen / AA

that $f_\phi < (\rho c_s h)\Omega \approx \rho c_s^2 \approx p$. 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, $p_{\rm rad} = 1/3 a_{\rm rad} T^4$. We define the fractional contribution of the gas to the total pressure as

$$\beta \equiv \frac{p_{\rm gas}}{p}. \tag{7.14}$$

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 pressure or just the gas pressure, respectively.

Since the energy of each photon is just its momentum times the speed of light, the radiative energy flux is simply given by the change in the radiation pressure (momentum flux) per photon mean free path,

$$F = -c\,\frac{dp_{\rm rad}}{d\tau}, \tag{7.15}$$

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

$$\tau = \int_0^h \kappa\rho\, dz \approx \frac{1}{2}\kappa\Sigma, \tag{7.16}$$

where $\Sigma = 2h\rho$. For the characteristic mass density ρ and temperature T encountered at the midplane of accretion disks around supermassive black holes, there are two primary sources of opacity: *electron scattering* with

$$\kappa_{\rm es} = \frac{\sigma_{\rm T}}{m_p} = 0.4\ {\rm cm^2\ g^{-1}}, \tag{7.17}$$

and *free-free* absorption with

$$\kappa_{\rm ff} \approx 8 \times 10^{22} {\rm cm^2\ g^{-1}} \left(\frac{\rho}{\rm g\,cm^{-3}}\right)\left(\frac{T}{\rm K}\right)^{-7/2}, \tag{7.18}$$

where we have assumed a pure hydrogen plasma for simplicity.

change subscript from `Edd' to `E' \AA

It is customary to normalize the accretion rate $\dot{M}$ in the disk relative to the so-called Eddington rate $\dot{M}_E$, which produces the maximum possible disk luminosity, $L_{\rm Edd}$ (see the derivation in equation 7.33). When the luminosity approaches the Eddington limit, the disk bloats and h approaches r, violating the thin-disk assumption. We write $\dot{m} = (\dot{M}/\dot{M}_{\rm Edd})$, with $\dot{M}_{\rm Edd} \equiv (L_{\rm Edd}/\epsilon c^2)$, where ϵ is the radiative efficiency for converting rest mass to radiation near the ISCO (shown in Figure 7.5).

In local thermodynamic equilibrium, the emergent flux from the surface of the disk (equation 7.11) can be written in terms of the midplane disk temperature T as $F \approx c a_{\rm rad} T^4/\kappa\Sigma$. The surface temperature of the disk is then roughly

Insert paragraph from p. 229 here / AA

$$T_d \approx \left(\frac{4F}{a_{\rm rad}}\right)^{1/4} = 10^5\ {\rm K}\ M_8^{-1/4} \dot{m}_{-1}^{1/4} r_1^{-3/4} \left[1 - \left(\frac{r}{r_{\rm ISCO}}\right)^{1/2}\right]. \tag{7.19}$$

Note that the disk surface temperature increases at small black hole masses and reaches the X-ray regime for stellar-mass black holes. (Nonthermal X-ray emission from a hot corona or a jet can supplement this disk emission.) Stellar-mass black holes can therefore be important X-ray sources at high redshifts, especially if they get incorporated into a binary system where they accrete gas from a companion star. In the local Universe, black hole X-ray binaries come in two flavors, depending on the mass of the companion star: *low-mass X-ray binaries*, in which a low-mass companion transfers mass owing to the tidal force exerted by the black hole, and *high-mass X-ray binaries (BH-HMXBs* or *microquasars)*, in which the companion is a massive star that can also transfer mass to the black hole through a wind. At redshifts $z > 6$, when the age of the Universe was short, BH-HMXBs were probably most important, since they are known to produce their X-rays over a short lifetime ($< 10^9$ yr). The cumulative X-ray emission from BH-HMXBs is expected to be proportional to the star formation rate. If, indeed, the early population of stars was tilted toward high masses, and binaries were common, BH-HMXBs may have been more abundant per star formation rate in high-redshift galaxies. As we discuss elsewhere, the X-rays produced by BH-HMXBs may have had important observable effects as they catalyzed H_2 formation (§6.2), heated the IGM (§9.8.2), and modified the 21-cm signal from neutral hydrogen (§12.3.2). Their overall influence was, however, limited: hydrogen could not have been reionized by X-ray sources based on current limits on the unresolved component of the X-ray background. Throughout this chapter, we focus our attention on supermassive black holes, which are brighter and hence easier to detect individually at high redshifts.

Move to p. 228 at position marked

Utilizing the preceding equations, we are now in 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 $\epsilon_{-1} = (\epsilon/0.1)$.

e / PE

lowercase italics r subscript Roman Sch /AA

For supermassive black holes, the accretion disk can be divided radially into three distinct regions,[8]

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

$$r_{1,\rm im} \approx 54\,\alpha_{-1}^{2/21}(\dot{m}_{-1}/\epsilon_{-1})^{16/21}M_8^{2/21} \quad \text{if } b = 1, \tag{7.20}$$

$$58\,\alpha_{-1}^{2/21}(\dot{m}_{-1}/\epsilon_{-1})^{16/21}M_8^{2/21} \quad \text{if } b = 0, \tag{7.21}$$

and the transition radius between regions 2 and 3 is

$$r_{1,\rm mo} \approx 4 \times 10^2\,(\dot{m}_{-1}/\epsilon_{-1})^{2/3}. \tag{7.22}$$

The surface density and scale height of the disk are given by

Inner region:

$$\Sigma(r) \approx (3 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5} \quad \text{if } b = 1, \tag{7.23}$$

$$(8 \times 10^2\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-1}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{-1} r_1^{3/2} \quad \text{if } b = 0, \tag{7.24}$$

$$h(r) \approx R_{\mathrm{Sch}}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right). \tag{7.25}$$

lowercase italics
r subscript
Roman Sch /AA

Middle region:

$$\Sigma(r) \approx (3 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/5} M_8^{1/5} r_1^{-3/5}, \tag{7.26}$$

$$h(r) \approx 1.4 \times 10^{-2} R_S \alpha_{-1}^{-1/10}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{1/5} M_8^{-1/10} r_1^{21/20}. \tag{7.27}$$

Outer region:

lowercase italics
r subscript
Roman Sch /AA

$$\Sigma(r) \approx (6 \times 10^6\,\mathrm{g\,cm^{-2}})\alpha_{-1}^{-4/5}\left(\frac{\dot{m}_{-1}}{\epsilon_{0.1}}\right)^{7/10} M_8^{1/5} r_1^{-3/4}, \tag{7.28}$$

$$h(r) \approx 10^{-2} R_S \alpha_{-1}^{-1/10}\left(\frac{\dot{m}_{-1}}{\epsilon_{-1}}\right)^{3/20} M_8^{-1/10} r_1^{9/8}. \tag{7.29}$$

The midplane temperature is given by

$$T_m(r) \approx \left(16\pi^2\right)^{-1/5}\left(\frac{m_p}{k_B \sigma_T}\right)^{1/5} \alpha^{-1/5}\kappa^{1/5}\dot{M}^{2/5}\Omega^{3/5}\beta^{-(1/5)(b-1)}. \tag{7.30}$$

The preceding scaling laws ignore the self-gravity of the disk. This assumption is violated at large radii. The instability of the disk to gravitational fragmentation due to its self-gravity occurs when the so-called Toomre parameter, $Q = (c_s\Omega/\pi G\Sigma)$, drops below unity (see §5.2.3). For the preceding scaling laws of the outer disk, the onset of fragementation occurs at the outer radius,

e / PE

$$r_{1,\mathrm{out}} \approx 2 \times 10^4 \alpha_{-1}^{28/45}\left(\dot{m}_{-1}/\epsilon_{-1}\right)^{-22/45} M_8^{52/45}. \tag{7.31}$$

Outside this radius, the disk gas fragments into stars, and the stars may migrate inward as the gas accretes onto the black hole. The energy output from stellar winds and supernovae will supplement the viscous heating of the disk and may regulate the disk to have $Q \sim 1$ outside this 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 abundances of heavy elements in the spectra out to arbitrarily high redshifts. Since the total amount of mass in the disk interior to this radius constitutes 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.[9]

7.3.3 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 is 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 falls inward and its density ρ rises, the temperature T of each species 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 inward with increasing density as $T_e \propto \rho^{1/3}$, while the protons are nonrelativistic, with $\gamma = 5/3$, and so $T_p \propto \rho^{2/3}$. The result is a two-temperature plasma in which the protons are much hotter than the electrons. Typical analytic models[10] yield $T_p \sim 10^{12}\ {\rm K}(r/r_{\rm Sch})^{-1}$ and $T_e \sim \min(T_p, 10^{9-11}\ {\rm K})$. 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.

Italics / AA

Do not split / PE

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 energies 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). The thin-disk value applies to high accretion rates above some critical value, $\dot{m} > \dot{m}_{\rm crit}$, where $\dot{m}$ is the accretion rate (in Eddington units) near the ISCO, after taking into account the fact that some of the infalling mass at larger radii is lost to outflows. Analytic RIAF models typically admit of a radiative efficiency of

$$\frac{L}{\dot{M}c^2} \approx \epsilon \left(\frac{\dot{m}}{\dot{m}_{\rm crit}}\right) \tag{7.32}$$

for $\dot{m} < \dot{m}_{\rm crit}$, with $\dot{m}_{\rm crit} \sim 0.01$–$0.1$. For example, in the nucleus of the Milky Way, massive stars shed $\sim 10^{-3} M_\odot\ {\rm yr}^{-1}$ of mass into the radius of influence of central black hole Sgr A*, but only a tiny fraction $\sim 10^{-5}$ of this mass accretes onto the black hole.

Since mergers are rare at low redshifts and much of the gas in galaxies has already been consumed in making stars, most local supermassive black holes are characterized by a very low accretion rate. The resulting low luminosity of these dormant black holes, such as the $4 \times 10^6 M_\odot$ black hole lurking at the center of the Milky Way galaxy, is often described using RIAF/ADAF models. Although this mode of accretion is characterized by a low mass infall rate, it could persist over a period of time that is orders of magnitude longer than the

quasar mode discussed earlier, so its contribution to the growth of black holes in galactic nuclei may not be negligible.

7.4 The First Black Holes and Quasars

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, was derived in equation (5.27) by balancing the inward force of gravity on each proton by the outward radiation force on its companion electron (which is the momentum flux carried by the radiation times the scattering crossection of the electron) at a distance r:

$$\frac{GM_{\rm BH}m_p}{r^2} = \frac{L_E}{4\pi r^2 c}\sigma_T, \tag{7.33}$$

where m_p is the proton mass, and $\sigma_T = 0.67 \times 10^{-24}$ cm^2 is the electron cross section for scattering a photon (Thomson scattering). 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 equals the luminosity of massive stars that are held at rest against gravity by radiation pressure, as described by equation (7.33). This limit is formally valid in a spherical geometry, and exceptions to it were conjectured for other accretion geometries over the years (and we shall consider one such possibility for a spherical geometry in §7.4.1). But, remarkably, observed quasars for which black hole masses can be measured by independent methods appear to respect this limit. When we substitute all constants, the Eddington luminosity is given by

$$L_E = 1.3 \times 10^{44} \left(\frac{M_{\rm BH}}{10^6\ M_\odot}\right)\ {\rm erg\ s^{-1}}, \tag{7.34}$$

As discussed previously, we can write the total luminosity from gas accreting onto a black hole, L, as some radiative efficiency ϵ times the mass accretion

[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 behaves as a single quasi-neutral fluid that is subject to both forces. Under similar circumstances, electrons are confined to the Sun by an electric potential of about a kilovolt (corresponding to a total charge of $\sim$ 75 C). 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 kV 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).[11]

rate $\dot{M}$,

$$L = \epsilon \dot{M} c^2, \tag{7.35}$$

where the black hole accretes the nonradiated component, $\dot{M}_{\rm BH} = (1-\epsilon)\dot{M}$. The equation that governs the growth of the black hole mass is then

$$\dot{M}_{\rm BH} = \frac{M_{\rm BH}}{t_E}, \tag{7.36}$$

where (after we substitute all fundamental constants),

$$t_E = 4 \times 10^7 \left(\frac{\epsilon/(1-\epsilon)}{0.1}\right)\left(\frac{L}{L_E}\right)^{-1} \ {\rm yr}. \tag{7.37}$$

We therefore find that as long as fuel is amply supplied, the black hole mass grows exponentially in time, $M_{\rm BH} \propto \exp\{t/t_E\}$, with an e-folding time t_E. Since the growth time in equation (7.37) 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 was 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 to $\sim 10^9\ M_\odot$ for $\epsilon \sim 0.1$ and $L \sim L_E$. However, the intervention of various processes makes it unlikely that a stellar mass seed is able to accrete continuously at its Eddington limit without interruption.

Spacing should be same on either side of ~ / PE

For example, mergers were very common in the early Universe. Whenever two gas-rich galaxies come together, their black holes are likely to coalesce. The coalescence is initially triggered by "dynamical friction" from the surrounding gas and stars and is completed—once the binary gets tight—as a result of the emission of gravitational radiation.[12] 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 tightens 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 owing to the nonisotropic emission of gravitational radiation at the final plunge (see §7.7). This kick was calculated recently using advanced computer codes that solve Einstein's equations (a task that was plagued for decades with numerical instabilities).[13] The typical kick velocity is hundreds of kilometer per second (and up to ten times more for special spin orientations), much larger than the escape speed from the first dwarf galaxies. Thus, continuous accretion was likely punctuated by black hole ejection events[14] that forced the merged dwarf galaxy to grow a new black hole seed from scratch.

s / AA

DES: 2 lines under head ok?

7.4.1 Supermassive Stars

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*, defined as hydrostatic configurations with masses of 10^3 to $10^8\ M_\odot$. Such systems have not been observed yet. Theoretically, they are expected to be supported almost entirely by radiation pressure and hence their luminosity equals the Eddington limit, $L = 1.3 \times 10^{44} (M_\star / 10^6\ M_\odot)$ erg s^{-1}. Supermassive stars steadily contract and convert their gravitational binding energy to radiation with a total lifetime $< 10^6$ yr before they collapse to a black hole.

Add subscript `E' /AA

Was it possible to make such stars in early galaxies? Yes: numerical simulations indicate that stars of mass $\sim 10^6\ M_\odot$ could have formed at the centers of early dwarf galaxies that were barely able to cool their gas through transitions of atomic hydrogen and had $T_{\rm vir} \sim 10^4$ K and no H_2 molecules (which were suppressed by a Lyman-Werner background). Such systems had a total mass several orders of magnitude higher than the earliest Jeans-mass condensations discussed in §3.2. In both cases, the gas lacks the ability to cool well below $T_{\rm vir}$, and so it fragments into one or two major clumps. The simulation shown in Figure 7.6 results in clumps of several million solar masses, which inevitably form massive black holes. The existence of such massive seeds would have given a jump start to the black hole growth process.

First, we show that the envelope of such stars must be convective. The condition for convective instability is that the star exhibit a negative entropy gradient. This follows from the fact that convective eddies that are hotter and rarefied relative to their environment tend to rise toward the star's surface and decrease their density adiabatically (at constant entropy) in pressure equilibrium with their environment. If the background entropy decreases as the eddies rise, then they become even more rarefied relative to their environment (lower ambient entropy at the same pressure implies higher ambient density) and continue to rise even further, hence leading to an instability. The energy transport by convective eddies drives the star to a state of marginal stability, namely, nearly uniform entropy. Let us first show that in the absence of convection, a supermassive star tends to develop a negative entropy gradient as it radiates away its energy.

The entropy of each electron–proton fluid element in a supermassive star changes according to the local radiative heat flux $\mathbf{F}$ at a rate

$$\frac{2T_g}{m_p}\frac{\partial s}{\partial t} = -\frac{1}{\rho}\nabla \cdot \mathbf{F}, \tag{7.38}$$

where ρ, T_g, and s are the mass density, temperature, and specific entropy of the element, respectively. If the opacity is dominated by Thomson scattering, then the local radiative heat flux $\mathbf{F}$ is related to the radiation pressure gradient by

$$\mathbf{F} = \frac{-m_p}{\sigma_T}\frac{1}{\rho}\nabla p_{\rm rad}. \tag{7.39}$$

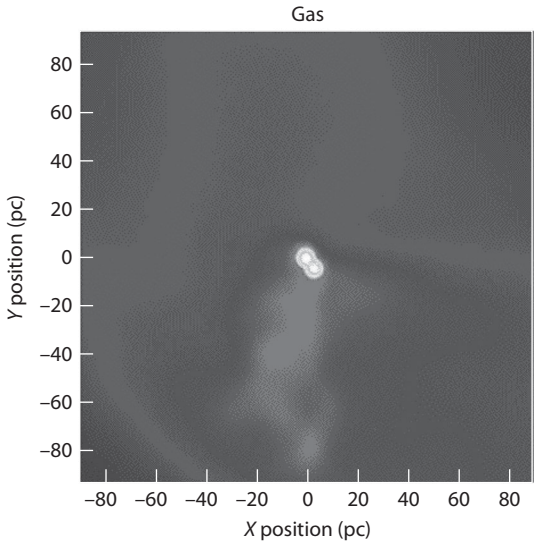


Figure 7.6 Numerical simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2 (see *Color Plate 14* for a color version of this figure). The image shows a snapshot of the gas density distribution 500 million years after 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 for supermassive black holes. The simulated box size is 200 pc on a side. Bromm, V., & Loeb, A., *Astrophys. J.* **596**, 34 (2003). Reproduced with permission of the American Astronomical Society.

by / AAS. /

If we ignore gas pressure and rotation, the hydrostatic equilibrium equation is simply

$$\frac{1}{\rho}\nabla p_{\rm rad} = \mathbf{g}, \tag{7.40}$$

where the gravitational field $\mathbf{g}$ obeys Poisson's equation,

$$\nabla \cdot \mathbf{g} = -4\pi G\rho. \tag{7.41}$$

Combining equations (7.39)–(7.41), we find that the right-hand side of equation (7.38) is constant,

$$\frac{1}{\rho}\nabla \cdot \mathbf{F} = \frac{4\pi G m_p}{\sigma_T} = \text{constant}. \tag{7.42}$$

Therefore, the gradient of equation (7.38) gives

$$\frac{\partial}{\partial t}\nabla s = \frac{2\pi G m_p^2}{\sigma_T}\frac{\nabla T_g}{T_g^2} < 0. \tag{7.43}$$

The radial temperature gradient is negative, since heat flows out of the star, which implies that the star will develop a negative entropy gradient and become convectively unstable. This result holds also for a rotating star, as long as the rotation period is much longer than the dynamical time, $t_{\rm dyn} \sim (G\rho)^{-1/2} = 1.1\ {\rm hr}(\rho/1\ {\rm g\ cm^{-3}})^{-1/2}$.

The nearly uniform entropy established by convection makes the structure of supermassive stars simple (equivalent to a so-called polytrope with an index $n = 3$) with a unique relation between their central temperature T_c and central density ρ_c,[15]

$$T_c = 2 \times 10^6\ {\rm K}\left(\frac{\rho_c}{1\ {\rm g\ cm^{-3}}}\right)^{1/3}\left(\frac{M_\star}{10^6\ M_\odot}\right)^{1/6}. \tag{7.44}$$

Because of this modest temperature, nuclear reactions are insignificant in metal-poor stars with masses $M_\star > 10^5\ M_\odot$. General-relativistic corrections make the star unstable to direct collapse to a black hole as soon as its radius contracts to a value

$$R_\star < R_{\rm crit} = 1.59 \times 10^3\left(\frac{M_\star}{10^6\ M_\odot}\right)^{1/2}\left(\frac{GM_\star}{c^2}\right). \tag{7.45}$$

Rotation can stabilize supermassive stars to smaller radii, but even rotating stars are expected to eventually collapse to a black hole after shedding their angular momentum through a wind. If the supermassive star is made of preenriched gas, then powerful winds will inevitably be driven at its surface, where the opacity due to lines from heavy elements far exceeds the Thomson value, making the outward radiation force stronger than gravity (see §7.5).

We note that the infall of a sufficiently dense, optically thick spherical envelope of gas cannot be prevented by radiation pressure even if the radiation production rate exceeds the Eddington limit near the center. To see this, let us consider a gas shell falling inward with a velocity $v_{\rm in}(r)$ at a radius r. If the outward diffusion time of photons through the gas, $t_{\rm diff} \sim \tau r/c$ (where $\tau \sim \sigma_T \rho r/m_p$ is the shell's optical depth to Thomson scattering), exceeds the infall time, $t_{\rm in} \sim r/v_{\rm in}$, then the radiation will be dragged by the infalling gas into the black hole. Even though the radiation is diffusing outward in the local rest frame of the gas, it is moving inward in the black hole frame of reference when $t_{\rm diff} > t_{\rm in}$. In that regime, the radiation will never be able to counteract the collapse of gas shells that are farther out. Expressing the mass accretion rate as

$$\dot{M} = 4\pi\rho r^2 v_{\rm in}, \tag{7.46}$$

we find that $t_{\rm diff} > t_{\rm in}$ if

$$\frac{\dot{M}}{\dot{M}_{\rm E}} > \epsilon\left(\frac{r}{GM/c^2}\right), \tag{7.47}$$

where M is the mass interior to radius r, and $\dot{M}_{\rm E} = L_{\rm E}/\epsilon c^2$ is the mass accretion rate that produces the Eddington luminosity (equation 7.34) for a radiative efficiency ϵ. We therefore conclude that as long as the mass infall rate is sufficiently high, the Eddington limit will not apply because of photon trapping. Super-Eddington accretion can therefore grow a seed black hole rapidly, as long as the blanket of infalling gas advects the radiation inward as it accretes onto the black hole. This "obscured" mode of black hole accretion (which is hidden from the view of observers) could be particularly important at high redshifts when the gas density and infall rate onto galaxies ottain their highest values.

7.5 Black Holes and Galaxies

7.5.1 The Observed Correlations between Supermassive Black Holes and Host Galaxies

As described in §7.1, there are clear correlations between the supermassive black holes that reside at the centers of galaxies and their host spheroids of stars. In particular, Figure 7.7 shows the strong correlation between the mass of the central black hole and the velocity dispersion of stars in the host galaxy's spheroid, $\sigma_\star$. (Note that this is *not* necessarily identical with virial velocity of the galaxy, because it samples motions on much smaller scales.) The data at $z \sim 0$ follow a tight correlation, $M_{\rm BH} \propto \sigma_\star^4$, with only a small apparent intrinsic scatter. Other correlations also exist; especially important is the observed relationship between the black hole mass and spheroid luminosity at $z \sim 0$, $M_{\rm BH} \propto L_{\rm sp}$, and that between the black hole mass and spheroid mass, $M_{\rm BH} \sim 0.0014 M_{\rm sp}$.[16]

These relationships explain how quasars may shine much brighter than their host galaxies. A typical star like the Sun emits a luminosity $L_\odot = 4 \times 10^{33}$ erg s^{-1}, which can also be written as a fraction $\sim 3 \times 10^{-5}$ of its Eddington luminosity $L_E = 1.4 \times 10^{38}$ erg s^{-1}. At the present day, 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, one to two orders of magnitude. The factor is smaller during short starburst episodes, which are dominated by massive stars with larger Eddington fractions.

Measuring black hole masses and host galaxy properties at high redshift is difficult, so the extrapolation of these relationships toward the cosmic dawn is not yet clear. The best way to do so is to resolve directly the region of influence of the black hole, where its gravitational potential dominates the motion of stars. Crudely, its mass can then be measured by fitting a Keplerian rise to the velocity dispersion at decreasing separations from the center. In practice, this is difficult because of the complex structure of galaxies and the projection of stars outside the region into the measured region. Nevertheless, the technique has been used successfully in dozens of nearby galaxies, as shown in Figure 7.7.

At higher redshifts, such direct observations are impossible, so indirect techniques are necessary. These rely on local "virial relations" between black hole

the / PE

change to 1.3 /AA

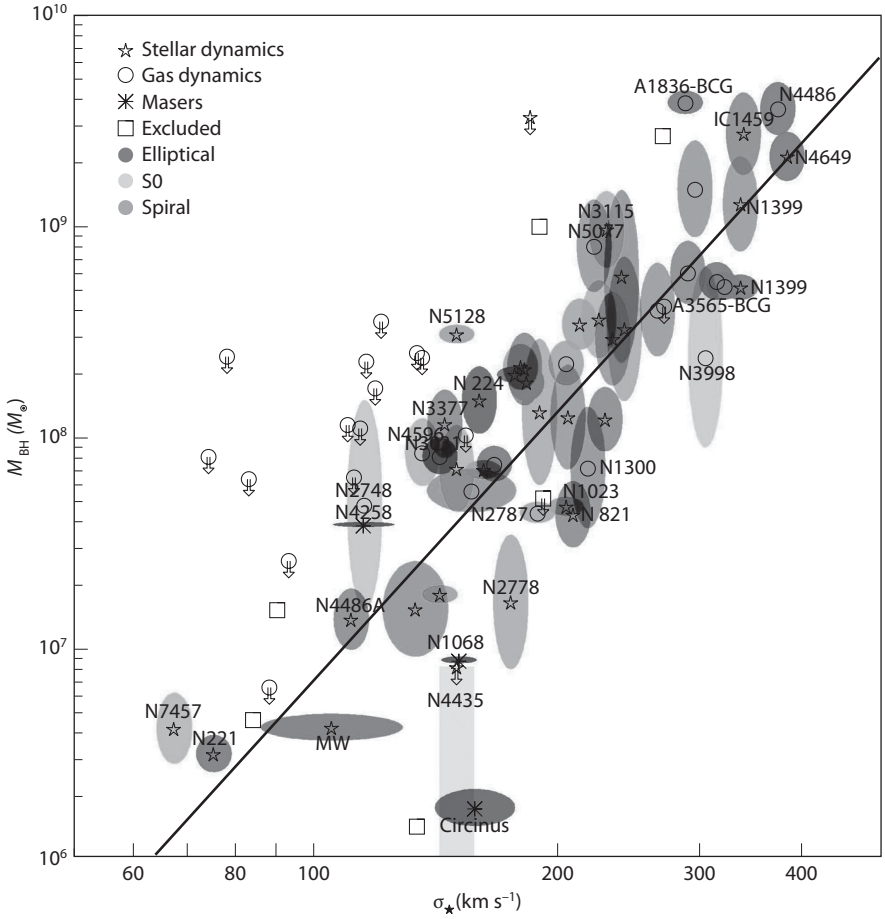


Figure 7.7 Dynamical measurements of the correlation between supermassive black hole mass, M_{BH}, and velocity dispersion of stars in the spheroid of its host galaxy, $\sigma_\star$ (see *Color Plate 15* for a color version of this figure). The symbol indicates the method of black hole mass measurement: dynamics of stars (*pentagrams*), dynamics of gas (*circles*), dynamics of maser sites (*asterisks*). Arrows indicate 3σ upper limits to black hole mass. The shade of the error ellipse indicates the Hubble type of the host galaxy: elliptical, S0, or spiral. The line is the best-fit relation to the full sample: $M_{BH} = 10^{8.12}\ M_\odot(\sigma_\star/200\ \text{km s}^{-1})^{4.24}$. The mass uncertainty for NGC 4258 has been plotted much larger than its actual value ~~so~~ show on this plot. Gültekin, K., et al., *Astrophys. J.* **698**, 198 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

to / PE

by / AAS. /

mass and spectroscopic properties of the AGN. These luminous sources typically show broad emission lines that (based on atomic physics considerations) are believed to originate from dense, highly ionized gas clouds near the black hole. Assuming that the velocity dispersion of these clouds v (which is of the order of a few thousand kilometers per second) is gravitationally induced by the

central black hole at their characteristic orbital radius r, we have $v^2 \sim GM_{\rm BH}/r$. The velocity can be measured from the Doppler width of the corresponding emission line. However, to obtain $M_{\rm BH}$ one also needs a measurement of the orbital radius of these clouds, which can be obtained through *reverberation mapping* observations that monitor the AGN spectrum over time. If the continuum emission (from the inner region of the accretion disk around the black hole) increases, then the broad-line region emission should also increase—but there will be a time delay as the continuum light propagates to that region, which can be measured with intensive monitoring programs. Cosmic time dilation makes this measurement more difficult for massive black holes (which are the only ones that can be easily observed) at high redshift.

Instead, an indirect proxy for the orbital radius is usually used. Empirically, the broad-line regions of all AGN are characterized by roughly the same UV flux, $L/4\pi r^2 \sim$ const. A possible interpretation of this scaling is that the emission lines originate near the outermost boundary where dust sublimates (and the cloud gas is exposed to the central UV illumination), which determines a particular threshold for the UV flux.[17] Thus, $r \propto L^{1/2}$ (with a proportionality constant that is fixed by the geometry and kinematics of the region), and the radius r can simply be estimated from the AGN luminosity. Measurements of nearby AGN are used to calibrate the proportionality constant.[18] The local AGN samples focus on emission lines easily accessible to optical telescopes observing at $z \sim 0$ (such as Hβ); extension to higher redshifts requires rest-UV lines (such as C IV) that redshift into the optical range. The calibration is not trivial, but it provides at least an indirect route to estimating black hole masses at high redshifts.

These techniques allow an estimate of the black hole mass, but to understand the relationship between the black hole and its host galaxy, one must also measure that galaxy's properties. This is also very difficult, because the only quasars visible at high redshifts are themselves extremely luminous—and hence dramatically outshine their host galaxies. Extremely good angular resolution is required to distinguish the pointlike AGN from the surrounding stars and gas (which is easiest to achieve for quasars that are magnified by intervening gravitational lenses). This measurement is most feasible in the radio, where the molecular line widths also provide a reasonably good estimate for the velocity dispersion and mass of the central star-forming portion of the galaxy. Assuming that we can classify that region as a spheroid, the highest-redshift quasars appear to have much more massive black holes than the local relations would suggest, with $M_{\rm BH} \sim 0.02 M_{\rm sp}$. The data are so far sparse, but interestingly, there appears to be a much weaker correlation with $\sigma_\star$ than in the local Universe.[19]

delete slash (do not close up) / PE

7.5.2 Galaxies and Their Supermassive Black Holes

Regardless of their extension to higher redshifts, the tight correlations in the local Universe described in §7.5.1 suggest an intimate connection between the growth of galaxy spheroids (or bulges) and their central black holes. This

delete slash (do not close up) / PE

relation, in turn suggests that either feedback from the black hole affects the galaxy's gas and stars, or feedback from those stars affects the black holes—or both. Both cases suggest that the relationship is an extremely important one for galaxy formation and evolution.

Since the mass of a galaxy at a given redshift scales with its virial velocity as $M_g \propto V_c^3$ (see equation 3.31), the binding energy of galactic gas scales as $M_g V_c^2 \propto V_c^5$. Meanwhile, the momentum required to expel the gas from its host scales as $M_g V_c \propto V_c^4$. Both scalings are reasonably close to the observed correlation shown in Figure 7.7 and can be tuned to explain the observed relations by appealing to feedback from the black hole to shut off star formation in the galaxy.

More concretely, suppose that accretion onto a black hole launches a wind that couples to the galaxy's ISM, driving it out of the potential well. As an example, suppose that this feedback contains a total energy that is a fraction ϵ_w of the rest mass of the black hole. Using similar arguments to those leading to equation (6.26), we find that this energy can unbind the gas in the galaxy once

$$M_{\rm BH} \sim 2 \times 10^4 \left(\frac{0.05}{\lambda}\frac{0.1}{\epsilon_w}\right)\left(\frac{M_h}{10^{10}\ M_\odot}\right)^{5/3}\left(\frac{1+z}{10}\right)\ M_\odot \tag{7.48}$$

$$\sim 2 \times 10^4 \left(\frac{0.05}{\lambda}\frac{0.1}{\epsilon_w}\right)\left(\frac{V_c}{78\ {\rm km\ s^{-1}}}\right)^{5}\left(\frac{1+z}{10}\right)^{-3/2}\ M_\odot, \tag{7.49}$$

where we have written the limit in terms of both the halo mass and circular velocity for convenience. One can then imagine that because this same gas is ultimately the fuel source for the black hole, both it and the galaxy's stellar population will stop growing once it reaches this limiting size, establishing the tight underlying correlation. Importantly, this limiting size implies that black holes—like supernovae—are much more efficient at expelling gas from small halos than large ones, so black hole may also play a very significant role at high redshifts.

s / PE

Similarly, following the arguments leading to equation (6.31), we see that the momentum input from a black hole wind will suffice to drive the gas out of a galaxy if[ii]

$$M_{\rm BH} \sim 6 \times 10^6 \left(\frac{0.1}{\epsilon_w}\right)\left(\frac{M_h}{10^{10}\ M_\odot}\right)^{4/3}\left(\frac{1+z}{10}\right)\ M_\odot \tag{7.50}$$

$$\sim 6 \times 10^6 \left(\frac{0.1}{\epsilon_w}\right)\left(\frac{V_c}{78\ {\rm km\ s^{-1}}}\right)^{4}\left(\frac{1+z}{10}\right)^{-1}\ M_\odot. \tag{7.51}$$

As expected, this approach provides a more stringent limit because it assumes that the outflow loses energy as it propagates through the galaxy.

This general picture is an attractive one, but note that neither of these simple scenarios has a straightforward connection to the observed correlations. This

[ii]Here we use $V_c^2 = 2\sigma_\star^2$ for a singular isothermal sphere profile, and we assume that the black hole releases most of its energy over a dynamical time $\sim \sigma_{\rm vir}/r_\star$.

is easiest to see by noting that although the scaling with velocity could easily be *interpreted* as mimicking the observed scaling with $\sigma_\star$, that would require a nonlinear relationship between spheroid mass (or luminosity) and halo mass. A key question for understanding how black holes interact with their hosts is then obviously how the spheroid properties like $\sigma_\star$ relate to the halo properties. One simple correction may be to note that if the fuel supplies of stellar bulges and supermassive black holes are identical, a linear relation between bulge mass and black hole mass may be naturally described by the feedback-regulated growth picture described. In that case, the relation between bulge mass and velocity dispersion is an interesting one that reflects how galaxies accumulate and process cold gas.

This relationship then raises an interesting "chicken and egg" problem, because *both* the black hole and stellar component generate feedback on the gas supply, and it is difficult to diagnose from observations which is more significant. The energy injected by supernovae of bulge stars is $\sim \omega_{\rm SN} M_{\rm sp}$, where $M_{\rm sp}$ is the stellar mass of the bulge component, and $\omega_{\rm SN}$ is the energy output per unit mass in stars, $\sim 10^{49}$ erg $M_\odot^{-1}$ for typical stellar populations. The energy injected by the black hole is $\sim \epsilon_w f_{\rm BH} M_{\rm sp} c^2$, where $f_{\rm BH}$ is the fraction of the stellar bulge mass in the black hole ($\sim$0.0014 in the nearby Universe, and possibly increases in massive halos at higher redshift; see §7.5.1). Then, the energy output from the black hole dominates if

$$\epsilon_w > 0.004 \left(\frac{0.0014}{f_{\rm BH}}\right) \left(\frac{\omega_{\rm SN}}{10^{49}\ {\rm erg}\ M_\odot^{-1}}\right), \tag{7.52}$$

which is clearly very small. However, as we discuss in §7.5.3, it is not clear how this energy reservoir couples to the galaxy ISM.

The answers to these questions depend on the way in which black holes (and the central stellar regions) are fed by gas in their host galaxies, which requires either a way to dissipate the angular momentum of the gas or an external force that can torque the gas toward the center. Viscous dissipation (even from instabilities like the MRI described in §7.3.2) is quite slow in the ISM and is not thought sufficient to feed an AGN.

Instead, the most popular model is one in which supermassive black holes in galaxies are fed with gas in episodic events of gas accretion triggered by mergers of galaxies. Tides and gravitational interactions during major mergers (i.e., in which the two merging components have comparable mass) provide massive torques that efficiently drive a large fraction of the gas toward the center of the merging components. However, black hole growth does not *require* a merger; any other process that can drive gas toward the galaxy center will also suffice, such as global disk instabilities. These can themselves be triggered by accretion streams that feed the galaxy halo itself at a high rate.

In the merger case, the fundamental cosmological input to modeling the growth of black holes is the merger rate of dark matter halos. For example, it is possible to "dress up" the mass distribution of halos in Figure 3.10 using the excursion set model to estimate their merger rates (see §3.4.2). The key physical inputs are the black hole–halo mass relation and a light curve for the quasar.

The former can be fixed via the observed $M_{\rm BH}$–$\sigma_\star$ relation. The latter then determines the (observable) luminosity from the black hole properties. In its most crude form, this value requires a maximum luminosity (which can be set as a fixed fraction of the Eddington limit, for example) and a duty cycle (which may be related to the black hole growth time t_E of equation 7.37 or to the dynamical time of the host galaxy, which is approximately the merger timescale). This simple approach can be tuned to give good agreement with the data on the quasar luminosity function shown in Figure 7.1. In the crudest form of the model, reasonable agreement with observations requires that quasars deposit ∼5% of their Eddington luminosity in the ISM of their host galaxy.[20]

Regardless of its origin, the inflow of cold gas toward galaxy centers during the growth phase of their black holes is naturally accompanied by a burst of star formation. The fraction of gas not consumed by stars or ejected by supernova-driven winds continues to feed the black hole. It is therefore not surprising that quasars and starbursts coexist in ultraluminous galaxies, and that all quasars show strong spectral lines of heavy elements.[21] In fact, as discussed in §7.3.2, the outskirts of accretion disks are very likely gravitationally unstable and susceptible to fragmentation to form stars.

7.5.3 Jets and Winds: Black Hole Feedback Mechanisms

delete comma / PE

Although the general picture of black holes' exerting strong feedback on their host galaxies is very attractive, the details remain vague, primarily because little direct evidence exists that any given feedback mechanism affects the host galaxy. The coupling mechanism is unknown, though several plausible candidates do exist: the process could be related to either the bright radiation or fast outflows that are known to be produced by quasars.

One possibility traces back to the fact that the scattering cross section per unit mass for UV radiation on dust is larger than σ_T/m_p (which is used to set the Eddington luminosity in equation 7.34) by two orders of magnitude.[22] Although dust is destroyed within $\sim 10^4 G M_{\rm BH}/c^2$ by the strong illumination from an Eddington-limited quasar, it should survive at larger distances. Hence, the radiation pressure on dust would exceed the gravitational force toward 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 crosssection of a spectral line over frequency,

$$\int \sigma(\nu)\, d\nu = f_{\rm osc}\left(\frac{\pi e^2}{m_e c}\right), \tag{7.53}$$

delete comma / PE

where $f_{\rm osc}$ is the absorption oscillator strength, is typically orders of magnitude larger than the Thomson value. For example, the Lyman-α transition of hydrogen, for which $f_{\rm osc} = 0.416$ (see §4.2), provides an average cross section that 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.

A second possibility for launching outflows is a (magneto)hydrodynamical effect within the accretion disk. We have already described how radiatively inefficient accretion flows can launch outflows, but there is good evidence for outflows in strongly emitting AGN as well. These outflows take (at least) two types. The first are jets, as shown in Figure 7.2, which are thought to be launched from the inner regions of accretion disks owing to complex magnetohydrodynamical effects. The material inside travels relativistically and may carry a great deal of energy, comparable to the radiative luminosities of the sources. However, such jets have extremely narrow opening angles and do not appear to interact with much of the galaxy's ISM. They are therefore unlikely to provide efficient feedback. The second possibility is a more classic "wind" launched from the disk itself, perhaps along open magnetic field lines threading the disk.

However they are launched, there is good evidence for these winds in some quasars. Called *broad absorption line quasars* (BALQSOs), these objects have strong redshifted absorption lines with extremely large widths $v_{\rm BAL} \sim 0.1c$ that are thought to represent outflowing material. The total kinetic luminosity in these winds is difficult to measure:

$$L_K \approx 2\pi f_c N_{\rm H} m_p v_{\rm BAL}^3 R_{\rm BAL}, \tag{7.54}$$

where f_c is the covering factor of the outflow, $N_{\rm H}$ is the column density of material in the absorber, and $R_{\rm BAL}$ is its distance from the central source. The covering fraction is uncertain, but a "unified" quasar model in which BALQSOs differ from "normal" quasars only because of the observer's viewing angle implies that $f_c \sim 0.1$. The radius of the absorbing material is even more uncertain, relying on modeling of the gas or outflow shock. Models suggest that $\epsilon_w \sim 0.03$ for BALQSOs, though with a large uncertainty.[23]

Although a wind can clearly couple to the ISM better than a narrow jet, it is still not obvious how the wind propagates from the central AGN point source through the much larger star-forming region. A great deal of work is required to understand the physics of quasar outflows and their influence on their host galaxies.

7.5.4 Black Holes and "External" Feedback

So far, we have focused on the interplay of black holes and their host galaxies. However, the possibility that black holes can expel large fractions of the gas from their hosts also indicates that their feedback mechanisms may influence the surrounding IGM and halos. Energetically, equation (7.52) suggests that even a modest coupling between feedback energy and the ISM or IGM suffices to make AGN outflows important relative to the radiative or supernova feedback from their host. There is even some direct evidence for AGN feedback in the local Universe: many galaxy clusters contain hot bubbles of material fed by radio jets from their massive central galaxies. Clusters also provide the strongest indirect evidence for intergalactic AGN feedback, where AGN are often invoked to explain the lack of apparent cooling in the cluster gas. (More generally, models often appeal to them to avoid the "overcooling problem"

described in §8.2, though there is little direct evidence for their activity in dark matter halos smaller than groups of galaxies.)

Fortunately, these large-scale black hole outflows can be modeled in the same way as their stellar counterparts (see §6.5.1), because of the approximate self-similarity of blast waves in an expanding Universe. AGN outflows will have many of the same effects as stellar superwinds, though the details will, of course, differ. For example, quasar winds may not carry as much of the host ISM out of the galaxy as their stellar counterparts and will therefore cause less IGM metal enrichment. However, plasmas launched from the highly magnetized accretion disk may drag magnetic fields with them, seeding an intergalactic magnetic field at high redshifts.

7.6 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 with the background gas and stars and ultimately coalesce through the emission of gravitational radiation.

When a tight binary is created 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 $M_{\rm halo}$ will sink to the center in 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 is reduced by a factor of [24] $\sim \varepsilon^{0.4}$. Therefore, massive satellites with $M_{\rm sat} > 0.1 M_{\rm halo}$ bring their supermassive black holes to the center of their host halos over 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 which the mean mass density of the satellite is comparable to the ambient density of the host galaxy.[25] 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 is typically expelled from the galactic nucleus at a high speed. This expulsion occurs at the so-called hardening radius of the binary,

$$a_{\rm hard} \approx 0.1 \frac{q}{(1+q)^2} M_6 \left(\frac{\sigma_\star}{100 \ {\rm km \ s^{-1}}} \right)^{-2} \ {\rm pc}, \qquad (7.55)$$

at which the binding energy per unit mass of the binary exceeds $3/2\sigma_\star^2$, where $\sigma_\star$ is the velocity dispersion of the stars before the binary tightened. Here, $M \equiv (M_1 + M_2)$ is the total mass, M_1 and M_2 are the masses of the two black holes,

insert parentheses / PE

$M_6 = (M/10^6 M_\odot)$, $q = M_1/M_2$ is the mass ratio. We also let $\mu = M_1 M_2/(M_1 + M_2)$ be the reduced mass of the binary.

A hard binary will continue to tighten only by expelling stars that cross its orbit. Unless the lost stars are replenished by new stars scattered into an orbit that crosses the binary (through dynamical relaxation processes in the surrounding galaxy, whose relaxation time is typically very long), the binary will 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 $a_i = (\mu/M_i)a$, where $i = 1, 2$. 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

$$P = 2\pi (GM/a^3)^{-1/2} = 1.72 \times 10^{-2}\, a_{14}^{3/2} M_6^{-1/2}\ {\rm yr}, \tag{7.56}$$

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

$$v = v_1 + v_2 = (2\pi a/P) = 1.15 \times 10^4 M_6^{1/2} a_{14}^{-1/2}\ {\rm km\ s^{-1}}. \tag{7.57}$$

In gas-rich mergers, the rate of inspiral decreases as soon as the gas mass interior to the binary orbit falls below μ, when the enclosed gas mass is no longer sufficient to carry away the entire orbital angular momentum of the binary, J. Subsequently, momentum conservation requires that fresh gas steadily flow toward the binary orbit for it to shrink. The binary tightens by expelling gas out of a region twice as large as its orbit (similarly to a blender's opening a hollow gap) and by torquing the surrounding disk through spiral arms. Fresh gas reenters 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 $\sim va$, 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 black holes and fuels quasar activity, it is appropriate to express $\dot{M}$ in Eddington units, $\dot{m} \equiv \dot{M}/\dot{M}_E$, corresponding to the accretion rate required to power the limiting Eddington luminosity with a radiative efficiency of 10%, $\dot{M}_E = 0.023\, M_\odot\ {\rm yr}^{-1} M_6$. We then find that

$$t_{\rm gas} \approx (J/\dot{M} v a) = \mu/\dot{M} = 1.1 \times 10^7\, \zeta \dot{m}^{-1}\ {\rm yr}. \tag{7.58}$$

For a steady $\dot{m}$, the binary spends equal amounts of time inside each logarithmic interval of radius a until gravitational waves start to dominate its loss of angular momentum.

insert:
the gas-driven coalescence timescale is
\AA

The coalescence timescale due to gravitational wave emission is given by[26]

$$t_{\rm GW} = \frac{5}{256}\frac{c^5 a^4}{G^3 M^2 \mu} = 2.53 \times 10^3 \frac{a_{14}^4}{\zeta M_6^3}\ {\rm yr}. \tag{7.59}$$

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

$$v_{\rm GW} = 4.05 \times 10^3\ \zeta^{-1/4} (\dot{m} M_6)^{1/8}\ {\rm km\ s^{-1}}; \tag{7.60}$$

$$P_{\rm GW} = 0.4\ \zeta^{3/4} M_6^{5/8} \dot{m}^{-3/8}\ {\rm yr}; \tag{7.61}$$

$$a_{\rm GW} = 2.6 \times 10^{-4}\ \zeta^{1/2} M_6^{3/4} \dot{m}^{-1/4}\ {\rm pc}. \tag{7.62}$$

For a binary redshift z, the observed period is $(1 + z) P_{\rm GW}$. The orbital speed at which gravitational waves take over is very weakly dependent on the supply rate of gas, $v_{\rm GW} \propto \dot{M}^{1/8}$, and generically corresponds to an orbital separation of the order of $\sim 10^3$ R Sch. The probability of finding binaries deeper in the gravitational wave–dominated regime, $\mathcal{P} \propto t_{\rm GW}$, diminishes rapidly at increasing orbital speeds, with $\mathcal{P} \propto (v/v_{\rm GW})^{-8}$.

lowercase italics r subscript Roman Sch / PE and AA

Black hole binaries can be identified visually or spectroscopically. At large separations the cores of the merging galaxies can easily be 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, IR, or radio wavelengths. The UV illumination by a quasar usually produces narrow lines from gas clouds at kiloparsec distances within its host galaxy or broad lines from denser gas clouds at subparsec 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 kiloparsecs 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 multikiloparsec separation; *(iii)* between narrow lines and broad lines if the black hole binary separation is between the kilopersec and parsec scales.

The last of these offset signatures also can be produced by a single quasar that gets kicked out of the center of its host galaxy while carrying the broad-line region with it. Such a kick can be produced either by the anisotropic emission of gravitational waves during the coalescence of a binary (which ~~generators~~ a recoil of up to $\sim$200 km s^{-1} in a merger of nonspinning black holes, and up to $\sim$4,000 km s^{-1} for special spin orientations), or from triple black hole systems that form when a third black hole is added to a galaxy center before the binary there has coalesced.[27] 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. The expected kick velocities of these early recoils offer a very interesting possibility for fossil signatures in the local Universe, because they are much larger than the escape speeds of the dwarf galaxy hosts but comparable to or smaller than the escape speeds of these galaxies' descendants, like the Milky Way. For example, the hierarchical formation of the Milky Way may have left recoiled black holes floating in its

generates / PE

halo, which are detectable through the compact star clusters that remain bound to these intermediate-mass black holes following their ejection from their host dwarf galaxies at high redshifts.[28]

7.7 Gravitational Waves from Black Hole Mergers

As just described, the final phase of black hole binary coalescence is driven by the emission of gravitational waves. The emitted waves could be detected by new observatories that are currently being planned or constructed.

As long as the binary separation is much larger than its Schwarzschild radius, the emitted gravitational wave luminosity can be derived in the post-Newtonian approximation. For two black holes on a circular orbit, the luminosity is[29]

$$L_{\rm GW} = \frac{32}{5}\frac{G^4}{c^5}\frac{M^3\mu^2}{a^5}. \tag{7.63}$$

The loss of energy to the emitted waves leads to a decrease in the binary separation a and an eventual coalescence of the two black holes after a time $t_{\rm GW}$ given by equation (7.59). Supermassive binaries with comparable mass members merge in less than a Hubble time once their separation shrinks to $a < 10^{3.5} r_{\rm Sch}$ (see equation 7.2) or once their relative orbital velocity $v = (GM/a)^{1/2} > 10^{-2}c = 3 \times 10^3$ km s^{-1}.

Future detectors will be sensitive to the gravitational wave amplitude. To an order of magnitude, the observed wave amplitude from an equal-mass binary with a Schwarzschild radius $r_{\rm Sch}$ and an orbital velocity v is given by $h \sim (1+z)(r_{\rm Sch}/d_L)(v^2/c^2)$, where d_L is the luminosity distance to the binary.[30] Since the signal amplitude declines only as the inverse distance rather than the inverse distance squared (as for electromagnetic detectors that respond to photon flux), the first generation of sensitive gravitational wave observatories will be able to find sources at cosmological distances.

More accurately, in a reference frame centered on the solar system's barycenter, the gravitational wave amplitude in its two polarization states is given by[31]

$$h_+ = \frac{2\mathcal{M}_z^{5/3}[\pi f_{\rm obs}]^{2/3}}{d_L}\left[1 + (\hat{\mathbf{L}}\cdot\hat{\mathbf{n}})^2\right]\cos[2\Phi(t)]; \tag{7.64}$$

$$h_\times = -\frac{4\mathcal{M}_z^{5/3}[\pi f_{\rm obs}]^{2/3}(\hat{\mathbf{L}}\cdot\hat{\mathbf{n}})}{d_L}\sin[2\Phi(t)]; \tag{7.65}$$

should be multiplication symbol, not letter x / AA

where the so-called redshifted chirp mass $\mathcal{M}_z \equiv (1+z)\mu^{3/5}/M^{2/5}$ sets the rate at which the binary shrinks, determining the "chirp" of its orbital frequency $P = 2\pi/\sqrt{GM/a^3}$. The precise orbital phase of the binary $\Phi(t)$ then depends on the masses and spins of the binary members and yields the observed wave frequency $f_{\rm obs}(t) = [\pi]^{-1}(d\Phi/dt)$, which is $(1+z)$ times smaller than the emitted wave frequency. The unit vector $\hat{\mathbf{n}}$ points from the solar system frame to the binary—defining the sky coordinates of the source, and the unit vector $\hat{\mathbf{L}}$

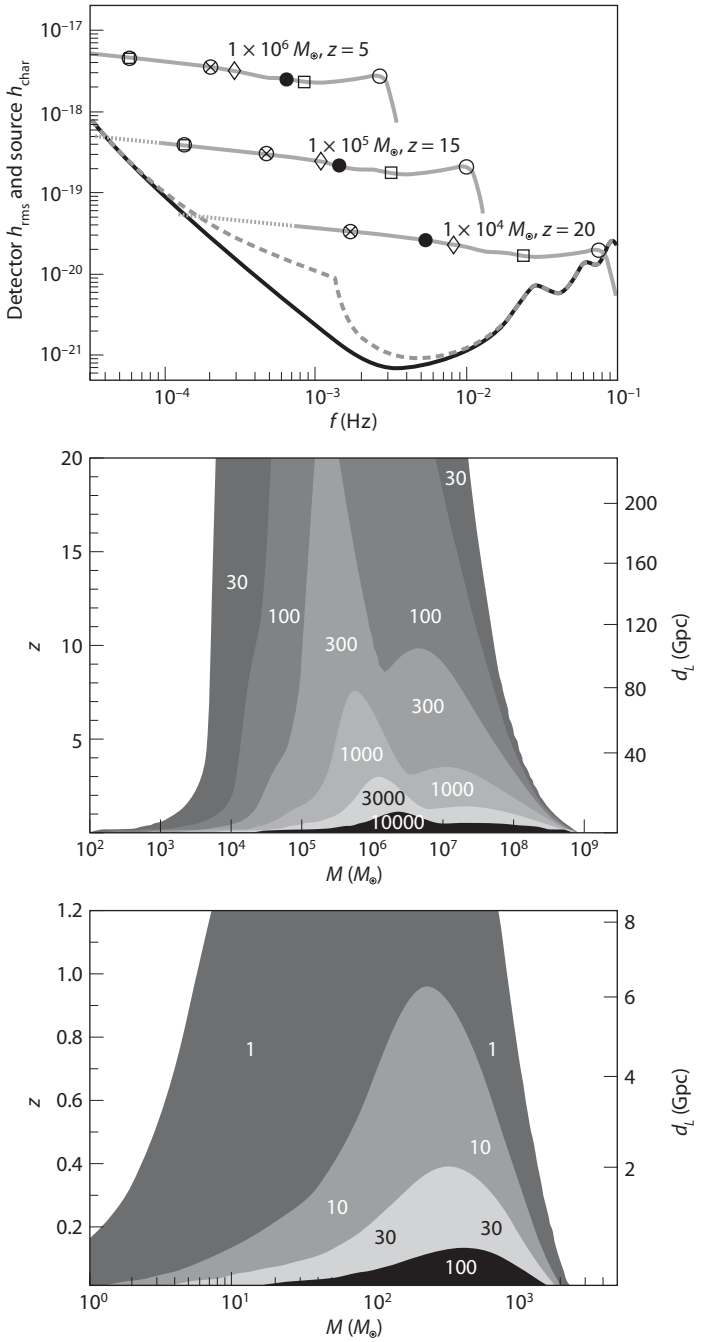


Figure 7.8 Sensitivity of the future gravitational wave observatories, LISA and Advanced-LIGO, to equal-mass ($M_1 = M_2 = M/2$) binaries (see *Color Plate 16* for a color version of this figure). *Top panel:* Root-mean-square noise

points along the direction of the binary angular momentum vector—defining the binary orientation relative to the line of sight. The inspiral signal does not explicitly provide the cosmological redshift separately from the binary masses, but the redshift can be inferred from $d_L(z)$ (or from an electromagnetic counterpart to the gravitational wave signal). Any particular detector is sensitive to a linear combination of the two polarization signals, with coefficients that depend on the orientation of the source relative to the detector.

Figure 7.8 shows the sensitivity of various gravitational wave observatories. The Laser Interferometer Space Antenna (LISA)[32] is a planned interferometer consisting of three spacecraft whose positions mark the vertices of an equilateral triangle 5 km on a side in an orbit around the Sun. As evident from Figure 7.8, LISA will be able to detect $\sim 10^4$ to 10^7 $M_\odot$ binaries out to arbitrary redshifts during the cosmic dawn. The next-generation ground-based interferometer, Advanced-LIGO,[33] will be sensitive to binaries involving black hole remnants of massive Population III stars (with $\sim 10^2$ to 10^3 $M_\odot$) out to $z \gg 1$.

An electromagnetic counterpart would greatly reduce the positional error and also determine the redshift of the gravitational wave source, enabling its use as a "standard siren" for precision cosmological distance measurements. Electromagnetic radiation could naturally be produced by a circumbinary accretion disk prior to coalescence, through a variety of possible effects: viscous dissipation of gravitational wave energy in the disk might result in a weak electromagnetic transient shortly after the merger, reequilibration of the inner edge of the disk could create an X-ray brightening on a timescale of 10–10^3 yr, and shocks produced by the remnant's recoil might generate electromagnetic reverberations that might take $\sim 10^4$ years to dissipate as enhanced IR luminosity.[34] It is not obvious whether these signals could be distinguished from the much more abundant sources of temporal variability in quasars with a single black hole. Moreover, the luminosity of any circumbinary disk is expected to be significantly reduced by the cavity associated with the decoupling of the black hole binary from the inner edge of the disk in the final stage of inspiral. The disk is not expected to refill the cavity and return to its full luminosity for at least a decade after coalescence. On longer timescales, the portion of the accretion

PE

Figure 7.8 (Continued) amplitude of LISA $h_{\rm rms}$ from the detector only (dashed) and from the detector combined with the anticipated foreground confusion (dash-dotted), along with the characteristic amplitudes $h_{\rm char}$ of three binary masses M (solid). The locations on each $h_{\rm char}$ curve correspond to the peak amplitude (circle), 1 hour before the peak (filled circle), 1 day before the peak (circle with inscribed cross), and 1 month before the peak (circle with inscribed square) in the observer frame, as well as times of $25r_{\rm Sch}/c$ (square) and $500r_{\rm Sch}/c$ (diamond) before the peak in the source frame. *Middle panel:* Contour plot of the signal-to-noise ratio (SNR) with binary mass and redshift dependence for LISA. *Bottom panel:* SNR contour plot with mass and redshift dependence for Advanced-LIGO. Baker, J., et al., *Phys. Rev.* **D75**, 4024 (2007). Copyright 2007 by the American Physical Society.

Reprinted Figures 16, 17, and 19 with permission from /

AU: Please supply page numbers; article ID is 124024, they do not use normal page numbers, just 124024-#. The article is 17 pages long, so the last one is 124024-17

disk that remains bound to the recoiled black hole remnant is expected to be detectable as a kinematically and eventually a spatially offset quasar, although its lifetime is limited by the supply of gas that can remain gravitationally bound to it.[35] Interestingly, the recoil of the remnant black hole allows it to frequently disrupt stars along its path and produce a prompt electromagnetic signal that does not depend on the prior existence of a gaseous disk in the vicinity of the merging binary.[36] Tidal disruption of stars can be observed out to cosmological distances in cases where a relativistic jet is produced.[37]

The expected event rate of massive binary mergers can be calculated based on the halo merger rate predicted by the excursion set formalism (§3.4.1) under various assumptions about the relation between the black hole and halo masses. For reasonable assumptions, LISA is expected to detect many cosmological events per year. The actual detection of these signals would open a new window into the Universe and enable us to trace the hierarchical assembly of black holes in galaxies throughout cosmic history. Moreover, because gravitational waves pass freely through all forms of matter, gravitational wave observatories might discover new populations of black hole binaries that are electromagnetically faint because of their modest mass relative to bright quasars, or because they are enshrouded by gas and dust.

Chapter Eight

Physics of Galaxy Evolution

Let us summarize briefly what we have learned in the previous chapters. According to the popular cold dark matter cosmological model, dwarf galaxies started to form when the Universe was only 100 million years old. Lacking heavy elements to cool the warm primordial gas left over from the Big Bang to lower temperatures, the gas could have fragmented only into relatively massive clumps that condensed to make the first stars. Computer simulations indicate that these first stars were much more massive than the Sun. These stars were efficient factories of ionizing radiation. Once they exhausted their nuclear fuel, some of them 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. Somewhere along the way—either as remnants of massive stars or through direct collapse—black holes formed, merged, and accreted gas until they grew to become bright quasars. The UV radiation emitted by all generations of stars (and quasars) eventually leaked into the IGM and ionized gas far outside the boundaries of individual galaxies.

The earliest dwarf galaxies eventually merged and made bigger galaxies. 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.

Thus, it is galaxies—distant ancestors of our own Milky Way—that formed the building blocks of large-scale structure during the reionization era (and likely most of the cosmic dawn). In this chapter we examine these objects in some detail from a largely theoretical perspective. Along the way, we must bear in mind that although the described progression of events is plausible, at this time it is only a conjecture in the minds of theorists that has not yet been confirmed by observational data. We discuss many of these efforts in Part III of this book, but for now we focus on the physics that drives the objects.

8.1 High-Redshift Galaxies

The most fundamental processes in galaxy formation are those that drive the cycle of baryons from the IGM, into the galactic ISM, into stars, and back into the ISM (and possibly the IGM) through supernovae or stellar winds. Buried within these steps are a multitude of feedback processes, from photoionization,

supernovae, and gas accretion onto a central black hole. A comprehensive understanding of the physical details of these feedback mechanisms is lacking. Nevertheless, we can at least identify the important processes that drive the evolution of galaxies. In this chapter, we briefly describe these ingredients, dedicating special attention to how they might affect models and observations at high redshifts.

The starting point for understanding the abundance, clustering, and other properties of galaxies is the dark matter halo distribution, $n(m)$. We wish to understand the mapping from halo mass to luminosity (in many different bands or lines), stellar mass, metallicity, star formation history, velocity distribution, and any other physical properties of interest. Of course, there is nothing to demand that this mapping be one-to-one, or even that these physical properties depend exclusively on halo mass, as they could also depend on the halo's larger-scale environment, mass accretion history, and so forth. The challenge of research on galaxy formation and evolution is to understand which factors are most important and how all of them interact to produce the objects we observe.

In the crudest representation, galaxies are machines that transform accreted material (whether acquired through slow accretion of diffuse gas or through rapid mergers) into stars and black holes. The crucial complication is feedback, which can both prevent gas from accreting onto a halo in the first place and expel material that is already present (preventing it from forming stars or providing potential fuel for later accretion episodes). Because this feedback is generated on the smallest scales (through stars or black holes), understanding galaxy evolution requires a model spanning a large range of physical scales and processes.

Theoretical astrophysicists examine many of these problems individually (and hence generally in isolation from one another). Their results inform coarser models—including both "semianalytic" approaches that rely on relatively simple models for the many processes involved, as well as numerical simulations (which rarely span the required dynamic range and so also contain simple prescriptions for at least some of the processes). For the sake of brevity, we follow the former approach and aim only to parameterize the important processes and suggest some intuition for the underlying physics. This is by no means a comprehensive treatment but should give a flavor for the "less" exotic processes that affect galaxy evolution even at the present day. The full suite of physical inputs is described in many other books and review articles (see Further Reading for related resources).

The very simplest model involves two free parameters: *(i)* the fraction of baryons converted into stars within a host halo, $f_\star$, and *(ii)* the duty cycle of vigorous star formation activity during which the host halo is luminous and the stars are formed, $f_{\rm duty}$. This model defines the star formation timescale $t_\star$ as the product of $f_{\rm duty}$ and cosmic time $t_H(z) \approx 2/3H(z)$ at the redshift of interest z. The star formation rate $\dot{M}_\star$ is then related to halo mass M_h as follows:

$$\dot{M}_\star(M_h) = \frac{f_\star \times (\Omega_b/\Omega_m) \times M_h}{t_\star}. \tag{8.1}$$

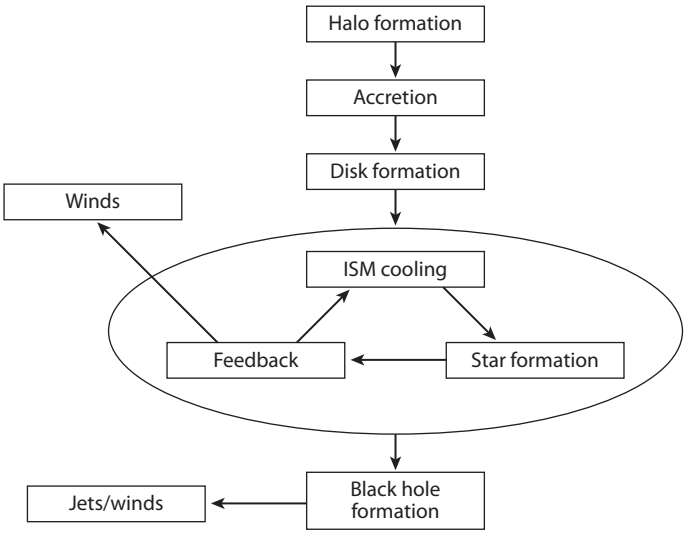


Figure 8.1 Diagram of the basic steps in galaxy formation, each of which is discussed in this chapter. The process begins with a dark matter halo that forms and accretes more matter (either through smooth accretion or mergers). The gas then settles into a rotationally supported "disk." Within the disk's ISM, self-regulation occurs through cooling and feedback, which ejects some fraction of the halo's gas through winds. Gas flowing toward the center of the galaxy may accrete onto a central black hole, which then unbinds other ISM gas through winds or jets.

Within the context of this simplest model, the physics of galaxies determines the values of $f_\star$ and $t_\star$ (which may depend on halo mass as well, and presumably have some scatter between different halos at the same mass). More complex models are obviously necessary to track more detailed properties of the sources. Some of these ingredients are described next; we summarize the major steps in Figure 8.1.

8.2 Gas Accretion

The fuel for star and black hole formation is provided by gas accretion onto the halo, either in a relatively slow, steady mode or in a stochastic "merger" mode. The first is relatively easy to describe and model (see Figure 8.2). First, consider a spherical system. Provided the gas accretes supersonically—as it will if the halo has a virial temperature $T_{\rm vir}$ larger than the IGM temperature—we generically expect an accretion shock to form, at a radius comparable to the virial radius of the halo. However, such a configuration is stable only if the hot gas behind the shock can support it. If, instead, this gas cools rapidly, the shock

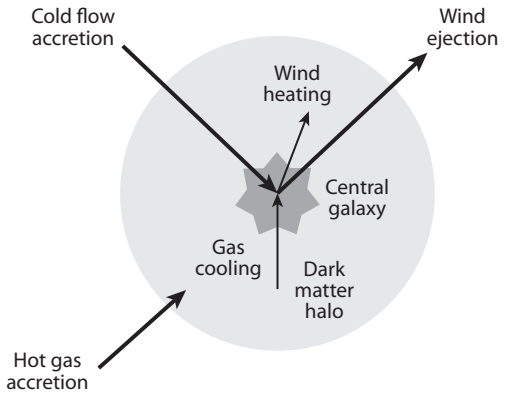


Figure 8.2 A diagram of the "baryon cycle" in which gas flows into and out of galaxies. The central galaxy is embedded in a much larger dark matter halo. Gas accreting onto the halo follows one of two fates: (1) in large halos, some is shocked at the virial radius and heated to high temperatures (from which point it must cool radiatively to settle onto the central galaxy); or (2) in small halos (and to some extent in larger ones), it travels along filamentary "cold flows" directly to the central galaxy. As the galaxy's gas is turned into stars, feedback launches winds, which both heat the halo gas (preventing it from cooling onto the galaxy) and expel some gas into the IGM.

will sink inward. The radiative cooling time is

Insert: energy /AA

$$t_{\rm cool} = \frac{3}{2}\frac{1}{\mu m_p}\frac{kT}{\rho(r)\Lambda(T,Z)}, \tag{8.2}$$

where $\rho^2\Lambda$ is the radiative cooling rate per unit volume. It is dominated by Compton cooling (at very high redshifts) and atomic line transitions, and therefore depends sensitively on the metallicity and temperature of the gas. Finally, the shock will stabilize very near the galaxy when the sound-crossing time $t_{\rm sc} \sim r/c_s$ becomes smaller than this cooling time. In more detail, the condition for virialization shock stability is[1]

$$\frac{\rho r \Lambda(T,Z)}{u^3 \mu^2 m_H^2} < 0.0126, \tag{8.3}$$

where T is the postshock temperature, and u is the infall velocity. The cooling rate therefore determines the geometry of the gas distribution inside the halo.

The transition between these rapid and slow cooling regimes depends on the halo mass and redshift. Crudely, halos with masses above a critical threshold $M_{\rm cool}$ have hot "atmospheres" that cool slowly. Gas accretes onto these atmospheres rather than the galaxy, and the rate of accretion onto the galaxy itself is limited not by cosmological processes but by cooling within the atmosphere. In contrast, halos with mass $M_h < M_{\rm cool}$ are limited only by the cosmological infall rate (and feedback from the galaxy itself; see further). This critical threshold occurs at $\sim 10^{11}\ M_\odot$ for gas with primordial composition, or at $\sim 10^{12}\ M_\odot$ for gas

with solar composition, with only a mild dependence on redshift (the left-hand side of equation 8.3 is $\propto \rho r_{\rm vir}\Lambda/u^3 \propto (1+z)^{1/2}\Lambda$). This mass is sufficiently large that most very high redshift galaxies are fed through the rapid-cooling regime and so are limited only by the rate of cosmological infall.

Recently, high-resolution simulations have shown that the filamentary geometry of the cosmic web changes this picture slightly: the "rapid-cooling" regime is generally fed by accretion along filaments that reaches the galaxy's star-forming region without any shocks until the gas strikes the high-density ISM. These *cold flows* provide the primary fuel supply for small galaxies, but the filamentary structures can persist in larger galaxies as well. The transition between these two cooling regimes is therefore not an abrupt one.

The overall growth of halo mass can be tracked with analytic arguments and simulations. In particular, the extended Press-Schechter formalism described in §3.4.2 provides a mechanism to estimate this growth, which matches numerical simulations reasonably well. In the standard cosmology, this approach yields[2]

$$\frac{\dot{M}_h/M_h}{H(z)} \approx 2.3 \left(\frac{M}{10^{10}\ M_\odot}\right)^{0.15} \left(\frac{1+z}{7}\right)^{0.75}, \tag{8.4}$$

which illustrates how rapidly accretion occurs at these very high redshifts. (In absolute terms, a $10^{10}\ M_\odot$ halo at $z = 7$ accretes gas at a rate of $2.6\ M_\odot\ {\rm yr}^{-1}$.)

Before discussing the fate of the cold flow gas in the small halos most important for high redshifts, we briefly describe a long-standing problem for halos in the slow-cooling regime, $M > M_{\rm cool}$, that motivates much contemporary work on galaxy evolution. Although gas in this regime does indeed cool slowly, it is still relatively fast by cosmological standards. Thus, high-mass galaxies at low redshifts should still have accreted most of their baryons and formed stars from them. However, observations show an exponential decline in the number density of galaxies (with $L_\star$ comparable to the Milky Way luminosity) at mass scales well below the exponential cutoff in the halo mass function $n(m)$ at the present time (which occurs near galaxy cluster scales). Evidently, then, some mechanism must prevent gas in massive halos from *overcooling* onto their central galaxies. This mechanism is likely to be feedback, operating either within the halos themselves or in the surrounding gas.

8.3 Halo Mergers

A second channel for adding mass to a galaxy is through a merger with a nearby halo. These merger rates are also described (roughly) by the extended Press-Schechter formalism, and equation (8.4) implicitly includes such growth in the overall accounting. However, the dynamics of such mergers differ greatly from smooth accretion.

Mergers are often divided into two classes, depending on the mass ratio of the merging systems. Let m_1 and m_2 be the masses of the two systems, with $m_1 > m_2$. *Major mergers* are usually defined to have a mass ratio $m_1/m_2 < 4$. Such

an interaction is quite dramatic, and so-called violent relaxation (due to time-varying gravitational potentials during the interaction) largely determines the structure of the resulting merger remnant, which may bear no resemblance to the merging galaxies. (Indeed, the classical picture for the formation of elliptical galaxies is through major mergers of spirals.) ~~IM~~ *minor mergers*, the mass ratio $m_1/m_2 > 4$. The second system then imprints only a small perturbation on the first, and the remnant retains the overall structure of the more massive object.

roman In / PE

In particular, gravitational tides raised by mergers disrupt both the stars and gases inside the individual systems. The former can mix freely, but the latter collide via shocks, possibly triggering widespread star formation over short timescales (typically a few dynamical times of the interacting galaxies). Such *starbursts* can be closely studied in nearby galaxies and, indeed, often show evidence for strong gravitational perturbations. The rapid growth of high-redshift galaxies naively suggests that such starbursts may have been very common in the early universe. Mergers are difficult to model analytically, so they have most often been studied with numerical simulations. These show that equal mass mergers can enhance significantly the star formation efficiency over isolated systems when averaged over the merger time.[3] Mergers can also funnel gas toward the center of the remnant galaxy, fueling any supermassive black hole there and triggering an AGN phase (see §7.5).

8.4 Disk Formation

As halo gas cools, it loses the pressure supporting it against gravity and contracts to higher densities. This contraction continues until the gas becomes rotationally supported by its own angular momentum.

The net angular momentum J of a galaxy halo of mass M_h, virial radius $r_{\rm vir}$, and total energy E, is commonly quantified in terms of the dimensionless spin parameter

$$\lambda \equiv J|E|^{1/2}G^{-1}M_h^{-5/2}\,. \tag{8.5}$$

Expressing the halo rotation speed as $V_{\rm rot} \sim J/(M_h r_{\rm vir})$ and approximating $|E| \sim M_h V_c^2$ with $V_c^2 \sim GM_h/r_{\rm vir}$, we find that $\lambda \sim V_{\rm rot}/V_c$, that is, λ is roughly the ratio of the actual rotation speed to the maximal value above which it will break up.

After cooling, the gas settles to a rotationally supported disk. Let us write the disk mass as a fraction $\tilde{m}_d$ of the halo mass and let the disk angular momentum be a fraction $\tilde{j}_d$ of that of the halo. The size of the disk is set by rotational support. As a simple estimate, let us take an isothermal profile for the dark matter halo and neglect the self-gravity of the disk. We further assume that the disk has an exponential surface density profile,

$$\Sigma(R) = \Sigma_0 \exp(-R/R_d), \tag{8.6}$$

where R_d is the disk scale radius. The total disk mass is then $M_d = 2\pi\Sigma_0 R_d^2$. Because the circular velocity of an isothermal sphere is constant, the total

angular momentum of the disk is

$$J_d = 2\pi \int V_c \Sigma(R) R^2 \, dR = 2 M_d R_d V_c. \tag{8.7}$$

Setting this expression equal to a fraction $\tilde{j}_d$ of the total angular momentum of the halo, as in equation (8.5), we obtain an expression for the R_d [4]

$$R_d = \frac{1}{\sqrt{2}} \left(\frac{\tilde{j}_d}{\tilde{m}_d} \right) \lambda \, r_{\rm vir}. \tag{8.8}$$

Note that the factor $\tilde{j}_d / \tilde{m}_d$ is simply the specific angular momentum of the disk material. The characteristic density and scale height of an isothermal disk at the hydrogen temperature floor of $\sim 10^4$ K can then be easily derived as a function of redshift.[5] The assumptions behind this simple expression are questionable: the self-gravity of the disk likely cannot be ignored once it collapses to a small size; the dark matter profile is not exactly isothermal (and it may respond to the gravity of the disk as well); and finally, the disk may not have organized itself into a simple exponential profile. Nevertheless, it proves to be a useful model in comparison with simulations.

We also require some way to calibrate the specific angular momentum of the disk material and the spin parameter. The observed distribution of disk sizes in local galaxies suggests that the specific angular momentum of the disk is similar to that expected theoretically for dark matter halos, and so we assume that $\tilde{j}_d / \tilde{m}_d = 1$. The distribution of disk sizes is then determined by the distribution of spin parameters and halo masses.[6] Numerical simulations indicate that the former approximately follows a lognormal probability distribution,[7]

$$p(\lambda) \, d\lambda = \frac{1}{\sigma_\lambda \sqrt{2\pi}} \exp \left[-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2} \right] \frac{d\lambda}{\lambda} \tag{8.9}$$

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

Despite its potential flaws, this simple model shows the expected scaling of the disk sizes with redshift: the size of a disk at a fixed halo mass is expected to scale as $R_d \propto (1+z)^{-1}$. Observations do, indeed, indicate that the luminous cores of galaxies follow this expected trend over the wide redshift range $2 < z < 8$, as illustrated in Figure 8.3 (though note that these galaxies are binned by luminosity rather than mass).

For high-redshift galaxies, the primary lesson is that even though the angular diameter distance *decreases* with z at high redshifts, the small masses and rapid cooling of the halo gas likely mean that the sources are extremely compact. Figure 8.4 shows the extrapolated relation between galaxy size and redshift, calibrated by current data on the size distribution and luminosity function of high-redshift galaxies. The graph implies that even the James Webb Space Telescope (JWST) will be able to resolve galaxies at an AB magnitude limit $m_{\rm AB} < 31$ only out to a redshift of $z \sim 14$. The next generation of large ground-based telescopes will resolve all galaxies discovered with JWST, but only if they are sufficiently clumpy to enable detection above the bright thermal sky.

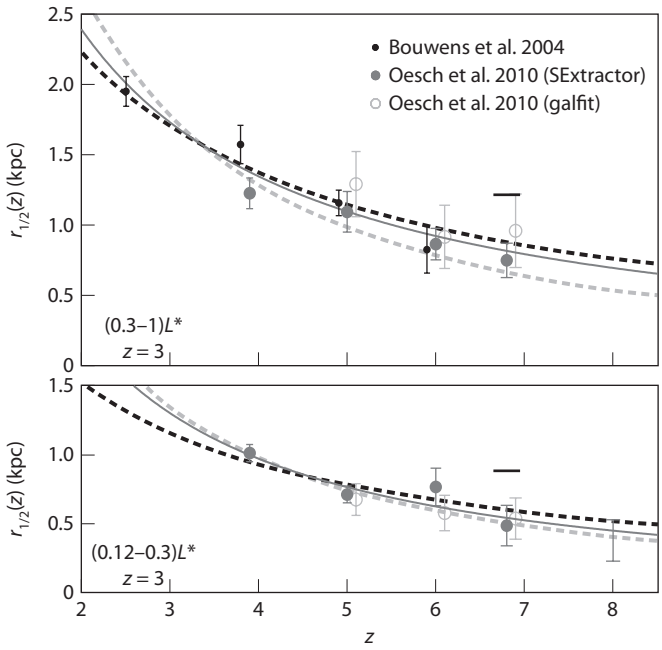


Figure 8.3 Observed evolution of the mean half-light radius of galaxies across the redshift range $2 < z < 8$ in two bins of fixed intrinsic luminosity: (0.3–1)$L_*(z = 3)$ (*top*) and (0.12–0.3)$L_*(z = 3)$ (*bottom*), where $L_*(z = 3)$ is the characteristic luminosity of a galaxy at $z = 3$ (equation 10.3). Different point types correspond to different methods of analyzing 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. Oesch, P. A., et al., *Astrophys. J.* **709**, L21 (2010). Reproduced with permission of the American Astronomical Society.

by / AAS. /

8.5 Star Formation in Galaxies

Once the gas has cooled and collapsed to high densities, star formation can commence. Determining the conversion efficiency of gas to stars is arguably the most important, and most challenging, aspect of galaxy formation. Nevertheless, theorists and observers have made enormous strides over the past few decades in understanding the relevant processes, at least in the local Universe. We illustrate these steps in Figure 8.5.

Traditionally, the star formation rate per unit area $\dot{\Sigma}_\star$ has been calibrated empirically as a function of the total gas surface density $\Sigma_{\rm gas}$. Observationally in the local Universe, these quantities correlate reasonably well over nearly seven

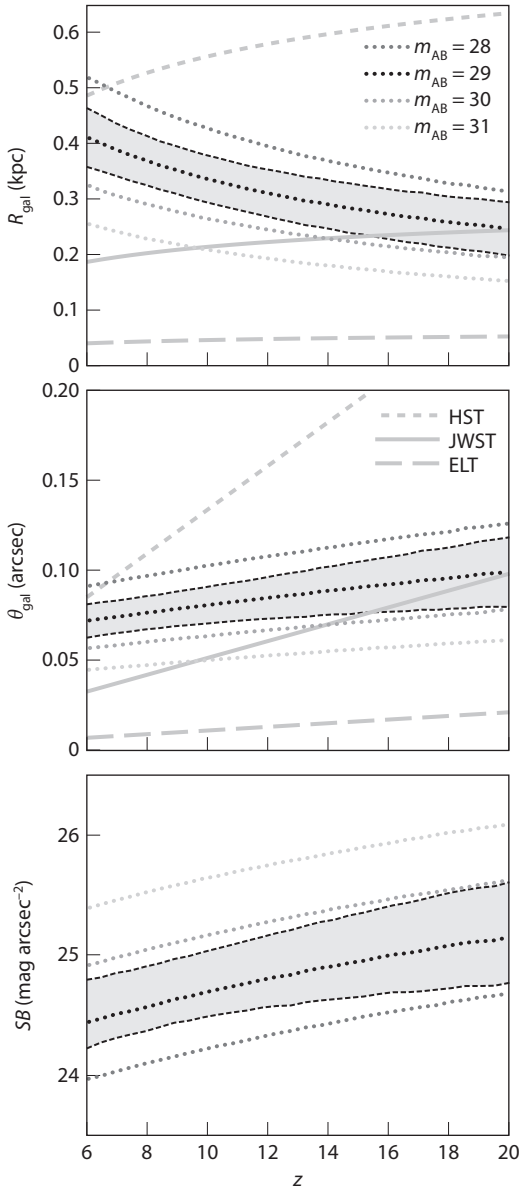


Figure 8.4 Theoretically extrapolated relation between galaxy size and redshift for four values of apparent AB magnitude. The *top* and *central* panels show the physical (R_{gal}) and apparent angular sizes (θ_{gal}), respectively. The thick gray lines indicate the resolution of telescopes with diameters corresponding to HST (2.5 m), JWST (6.5 m) and a ground-based extremely large telescope (ELT) (30 m). The *bottom* panel shows the average surface brightness within a galaxy scale radius as a function of redshift. In each panel, the gray band around the case of $m_{AB} = 29$ mag shows the 68% range of uncertainty on the mean. Wyithe, J. S. B., & Loeb, A., *Mon. Not. R. Astron. Soc.* **413**, L38 (2011). Copyright 2011 by the Royal Astronomical Society.

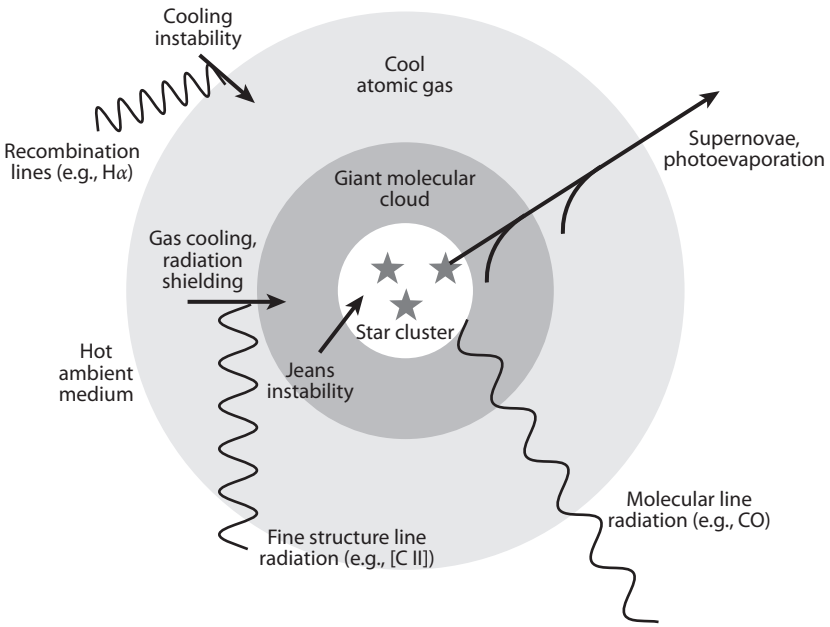


Figure 8.5 Diagram of baryon flows in the ISM. The multiphase ISM is a hot, ambient medium surrounding cold gas clouds, which form when a thermal instability allows the hot gas to cool locally, producing recombination line radiation (such as Hα). Inside these cool, neutral clouds, radiative cooling continues (largely through fine-structure lines such as [C II]), and the gas column eventually becomes sufficiently thick to shield H_2 from dissociating radiation, allowing a giant molecular cloud to form. Within the cloud, cooling continues via molecular transitions (such as CO), until regions became gravitationally unstable and form stars. The radiation and eventual explosions of these stars return gas and energy into the hot ISM, so that the process is self-regulating.

orders of magnitude in surface density, with

$$\dot{\Sigma}_\star \propto \Sigma_{\rm gas}^n, \tag{8.10}$$

where $n \approx 1.4 \pm 0.1$. This so-called Kennicutt-Schmidt relation[8] can also be interpreted in terms of the conversion of a fixed fraction of the gas into stars per orbital time in the associated galactic disks. In other words, if the (three-dimensional) star formation rate density satisfies $\dot{\rho}_\star \propto \rho_g / t_{\rm dyn}$, where ρ_g is the gas density, then we have $\dot{\rho}_\star \propto \rho_g^{3/2}$, which is very close to the observed two-dimensional relation. Despite the apparent success of this simple scaling, as an empirical relation it must still be tested beyond the local Universe, and it is unclear whether star formation would have obeyed the same relation at the low metallicity and low initial magnetization of the gas within the first galaxies.

Thus, a deeper understanding of star formation is highly desirable. As a first step, note that stars in the local Universe form in molecular clouds. One might therefore expect a more fundamental scaling of the star formation rate with

the density of *molecular* (rather than atomic) gas. We write f_{H_2} for the fraction of molecular gas. Furthermore, local observations show that molecular clouds turn a constant fraction $\epsilon_{\rm ff} \approx 0.015$ of their gas into stars per free-fall time.[9] This suggests a relation

$$\dot{\rho}_\star = \epsilon_{\rm ff} f_{H_2} \rho / t_{\rm ff} \tag{8.11}$$

for the star formation rate, which requires an estimate of the molecular fraction (and an extrapolation of the star formation efficiency parameter to high-z galaxies).

The molecular fraction is significantly more challenging to compute than the analogous calculation in §5.1, because enriched gas has more channels for H_2 formation (particularly on the surface of dust grains), a much more complex radiation field (owing to the embedded star formation), dust shielding, and a turbulent, inhomogeneous ISM. The physical picture that emerges is one in which molecular gas is confined to the interior of cold, high-density gas complexes. We must then determine *(i)* the relative mass of the cold phase and *(ii)* the fraction of the cloud able to go fully molecular. The latter is determined by balancing the rate of H_2 formation on dust grains with photodissociation by the (dust-extinguished) radiation field, using calculations similar to those presented in chapter 5.

The fraction of gas in the cold phase is determined ultimately by the feedback from hot stars and supernovae. The canonical picture assumes a multiphase ISM, with a "hot" phase of diffuse ionized gas and a cold phase of dense star-forming gas (and likely an intermediate warm phase of atomic gas). Crudely, gas is exchanged between the phases (as well as the stellar component) through three basic processes: *(i)* star formation (from cold gas to stars), *(ii)* cooling in the diffuse ISM (from hot to cold gas), and *(iii)* supernovae (from stars and cold gas to hot gas). The last process includes not only the supernova ejecta but also cloud evaporation through conduction with the surrounding hot gas.

Radiative cooling is challenging to model in the ISM unless the galaxy is fully resolved, because the density (and possibly composition) is highly inhomogeneous. For example, simply assuming a uniform cooling rate (even enhanced by a clumping factor) throughout the entire galaxy would not allow *any* gas to cool to very low temperatures. In reality, cooling is driven primarily by various thermal instabilities: because the cooling is most rapid in densest gas, this material quickly cools and becomes neutral, while the low-density gas remains hot. In other words, when cooling occurs, it is so efficient that a gas parcel usually drops out of the hot medium and rapidly enters the cold, molecular phase rather than mixing and lowering the temperature of the entire hot phase. In practice, the simple assumption of a two-phase medium, each with a characteristic *fixed* temperature, appears to provide a reasonable approximation. In this case, the radiated energy determines the mass flow rate from the hot to the cold phase.

Meanwhile, giant molecular clouds have much higher pressures than the surrounding ISM (at least in local galaxies), which suggests that their properties are set by internal feedback processes rather than by coupling to the ISM.[10] In particular, H II regions from embedded stars appear to provide the most

important feedback mechanism. Because they are internally regulated, the properties of these clouds do not vary much among galaxies, which explains the apparent constancy of $\epsilon_{\rm ff}$—though, of course, the conditions within high-redshift galaxies may be very different (for example, the ambient pressure will depend on its ISM density).

Numerical simulations have shown results consistent with equation (8.10), except that $n \approx 2$ for massive galaxies and $n \approx 4$ for dwarfs, as required by recent data.[11]

An alternative approach to this "bottom-up" view (which is fundamentally based on understanding the "microphysics" of star formation) is to treat it within a global context. The basic idea is that star formation can occur only if some sort of large-scale gravitational instability allows fragmentation to higher densities.[12] The condition for fragmentation to occur is the *Toomre criterion* (see §5.2.3),

$$Q \equiv \frac{c_s \kappa_e}{\pi G \Sigma_g} < 1, \tag{8.12}$$

where $\kappa_e = (2\Omega/R)\, d(\Omega R^2)/dR$ is the epicyclic frequency for an angular frequency of rotation $\Omega(R) = v/R$ at a (cylindrical) radius R within the disk. Once fragmentation begins, feedback from star formation heats the gas (and increases the sound speed), slowing further fragmentation. However, if star formation does not occur, the gas cools rapidly, decreasing Q. The expectation (which appears to be realized in nearby galaxies) is therefore that galaxies form stars sufficiently quickly to maintain $Q \sim 1$ in a sort of self-regulated flow. Together with a model for the feedback effects of stars, this provides an alternative method for determining $\dot{\Sigma}_\star$.

The two most obvious feedback mechanisms are stellar radiation pressure and supernovae, which "puff up" the disk and support the gas against the vertical component of gravity. Focusing on radiation pressure due to UV photons for concreteness, we can write (cf. the momentum injection rate from radiation in §6.4.1)

$$p_{\rm rad} \sim (1 - f_{\rm esc}) \epsilon \dot{\Sigma}_\star c, \tag{8.13}$$

where $\epsilon = 10^{-3}\epsilon_{-3}$ is the fraction of the baryonic rest energy converted to photons. (In nearby galaxies, supernovae produce a comparable pressure, but at high redshifts the elevated ambient densities make them less important.) The factor $(1 - f_{\rm esc})$ accounts for the fact that high-energy photons that escape the galaxy do not couple to the gas and so provide no radiation pressure.

As in §6.4.1, let us take the simple model of an isothermal density profile within the halo with a 1D velocity dispersion σ and assume that the disk contains a fraction f_g of the total matter. The fraction f_g is likely to be much larger than Ω_b/Ω_m, because the baryonic component has already cooled and collapsed into a disk. Assuming a thin disk, we can write the vertical component of hydrostatic equilibrium as $h \sim c_s/\Omega$, where Ω is the rotation rate (see §7.3.2). If we write $\Sigma_g = 2\rho h$, equation (8.12) provides an expression for the gas density inside the disk. With these two relations and $c_s^2 \sim p/\rho$, we can solve

equation (8.13) for the required star formation rate to support the disk:

$$\dot{\Sigma}_\star \sim 3 \left(\frac{Q}{\epsilon_{-3}(1-f_{\rm esc})}\right)\left(\frac{f_g}{0.25}\right)^2\left(\frac{\sigma}{50\ {\rm km\ s^{-1}}}\right)^4\left(\frac{100\ {\rm pc}}{r}\right)^2 M_\odot\ {\rm yr^{-1}\,kpc^{-2}}, \tag{8.14}$$

where we have scaled σ to a $10^{10}\ M_\odot$ galaxy at $z = 7$ (using $V_c^2 = 2\sigma^2$ for an isothermal sphere and equation 3.31).

This particular estimate ignores the pressure contribution from supernovae and a possible enhancement in the radiation pressure from IR emission by dust, but it gives a sense for how the global self-regulation criterion can be used to estimate the large-scale properties of galaxies. Such models typically connect more closely to the cosmological input parameters (the dark matter halo mass and its accretion rate $\dot{M}_h$). For example, given $\dot{M}_h$, one can integrate the star formation rate inward and determine the gas fraction f_g at each radius self-consistently.[13] The advantage of this approach is that it does not require a calibration to local galaxies and so is more robust to any unknown changes in the small-scale physics of star formation at high redshift; the disadvantage is that it makes strong assumptions about Q, the structure of the disk, and the relation between star formation feedback and the disk properties (ignoring other sources of pressure support like turbulence, for example).

However, the preceding relation does produce a surface-density law consistent with local models. If we define $\epsilon_{\rm SFR}$ via $\dot{\Sigma}_\star = \epsilon_{\rm SFR}\Sigma_g\Omega$, where Ω is the angular velocity (comparable to the dynamical time, and roughly the growth rate of global instabilities in disk galaxies), self-regulation at $Q \sim 1$ via radiation pressure yields $\epsilon_{\rm SFR}$ of a few percent for moderately large galaxies, with a predicted scaling $\epsilon_{\rm SFR} \propto \Sigma_g$.

8.6 Black Hole Growth in Galaxies

As described in chapter 7, it is now well established that nearly all present-day galaxies with spheroids also have supermassive black holes in their center. Because the properties of these black holes correlate with their host galaxies, and because their feedback may be important in regulating the stellar and gas contents of galaxies (see §7.5), it is natural to include them in models of galaxy formation and evolution.

Black holes may be fed smoothly and relatively slowly during the normal growth of a galaxy: some small fraction of the accreted gas may sink all the way through the galaxy and be swept into the black hole. The *minimal* accretion rate is given by the Bondi estimate from §5.2.1, $\dot{M}_{\rm BH} \sim G^2 M_{\rm BH}^2 \rho / c_s^3$, where the density ρ and sound speed c_s are evaluated at the accretion radius $R_{\rm acc} \sim G M_{\rm BH}/c_s^2$. However, this estimate generally produces slow accretion.

A more efficient method of feeding black holes is to channel gas toward them (see §7.5.2). This channeling can include any global instability (such as bars or spiral waves), but in the cosmological context galaxy mergers are often

identified as the most likely mechanism. As described previously, the torques generated during instabilities and mergers can be large, and a fair fraction of this gas can be fed toward the center of the remnant according to numerical simulations. Dimensionally, such strong torques produce inward radial velocities comparable to the local sound speed or orbital velocity,[14] many times larger than the expected effect of viscosity.

However, the fate of this gas is difficult to determine analytically, because it is of course also subject to star formation and feedback. Global disk models (as described in the previous section) can in principle follow the gas toward the galaxy's center, but a more phenomenological approach is often taken by assuming that the $M_{\rm BH}$–σ relation holds for all spheroidal galaxies and using it to *assign* a total accreted mass following a merger. As usual, one must worry about whether this relation holds during the earliest phases of galaxy formation (and, in particular, how it extrapolates with redshift; see §7.5.2).

Of course, if *both* galaxies in a merger have black holes, the resulting object will likely host a binary black hole. The fate of this merging system is described in §7.6, and it may have interesting signatures even beyond electromagnetic radiation. If the binary does not coalesce before the next merger, a triple (or higher multiple) system will form, from which the lightest black hole may be ejected at a speed of thousands of kilometers per second. Owing to the increase in the merger rate at high redshifts, multiple black hole systems are expected to have been more common in early galaxies.[15]

8.7 Feedback and Galaxy Evolution

Feedback from stars and black holes is crucial for galaxy evolution models in at least three respects. First, as we have already seen, it sets the properties of the star-forming gas within the galaxy itself, through radiation from stars and mechanical energy input via supernovae. Second, it enriches the gas, changing its dust content, cooling rates, and stellar properties. Finally, winds (whether driven by radiation pressure or supernovae) offer a second end point (other than stars) for accreted gas: it can be ejected entirely from the halo.

Modeling these different aspects is clearly very challenging, and often they are parameterized in simple fashions. For example, we have already seen that the "internal" feedback regulating star formation can be implicitly included in star formation laws with relatively simple phenomenological prescriptions like the Kennicutt-Schmidt law or its more recent modifications (though, again, one must always worry whether such prescriptions can be extrapolated robustly to the high redshifts of interest).

Chemical enrichment is probably the simplest of these effects to model: given an initial mass function, the rate at which material is enriched and returned to the ISM is straightforward to calculate. The ejecta are typically assumed to mix efficiently with the ISM, so that future generations of stars have monotonically increasing metallicity. The major uncertainties in chemical evolution are the properties of the gas accreting onto the galaxy (whether it is

preenriched) and the fraction of the ejected metals entrained into winds and carried out of the galaxy.

Perhaps the most significant aspect of feedback is mass loss through winds, which can dramatically affect the overall star formation efficiency in small galaxies. We have already discussed the complex physics of winds in §6.4. We expect feedback to be most important in the small gravitational potential wells of low-mass galaxies. The most crucial question is how the wind efficiency varies with galaxy mass, which depends largely on the underlying physics (i.e., momentum driven or energy driven). For example, suppose that the supernova energy accelerates a fraction of ISM material (with a mass loss rate $\dot{M}_w$) to the escape speed of the dark matter halo. Then, we have $\dot{M}_w \propto \dot{M}_\star \omega_{\rm SN}/v_{\rm esc}^2$. If, however, much of the energy is lost through cooling, so that the momentum input of the supernova or radiation pressure from the stars drives the wind, then $\dot{M}_w \propto \dot{M}_\star/v_{\rm esc}$, which has a significantly gentler dependence on $v_{\rm esc} \propto M_h^{1/3}$.

We note that beyond this overall scaling, the mass loading of the winds is also highly uncertain, because their total matter content is difficult to observe. So far, the best evidence comes from redshifted metal lines in galaxy spectra, which at best provide a velocity and column density of the material; without the distance of the absorbing material from the galaxy, the total mass is difficult to assess. Typically, however, the mass loss rate is assumed to be roughly equal to the star formation rate.

Finally, winds not only entrain gas but also can prevent circumgalactic gas from accreting onto the galaxy by heating it. This reduces the inflow rate onto the galaxy.

Given the basic energetics of the process, the prevailing expectation is that supernova feedback may suppress star formation in small galaxies and may help explain the relative dearth of low-luminosity galaxies compared with the number of small-mass halos. This hypothesis is consistent with local observations, where the total stellar mass is $\propto M_h^{2/3}$ for $M_h < M_{\rm crit} \sim 3 \times 10^{10}\ M_\odot$ and ~~constant~~ above it.[16] Assuming that the total supernova energy input $E_{\rm SN} \propto M_\star$, and that star formation continues until supernovae clear the halo of its remaining baryons by injecting an energy comparable to the binding energy of the gas, we would expect $M_\star/M_h \propto V_c^2 \propto M_h^{2/3}$. If this explanation applies at higher redshifts as well, we would expect a similar suppression there, in galaxies with $V_c < 100\,{\rm km\ s^{-1}}$.

proportional to italics M subscript italics h / AA

We should also consider feedback from black holes during accretion episodes of quasar activity, which may be important in driving galactic winds. As we described in §7.5, the energy input from quasars can exceed that from star formation, although the coupling of this energy to the ISM is not yet understood. (In some cases, such as jets from radio quasars, the energy may escape the galaxy in narrow channels without clearing all the gas.) However, the tight $M_{\rm BH}$–σ relation is highly suggestive of a fundamental relationship between the growth of black holes and their host galaxy's stars. Because (at least naively) this relationship suggests that the black hole mass scales superlinearly with halo mass, this feedback channel is *more* effective in larger galaxies and is often invoked as a potential solution to the "overcooling" problem in massive low-redshift galaxies.

Given the common "merger models" for AGN growth, one plausible physical picture is that the merger funnels large quantities of gas toward the remnant's center, triggering a starburst. Some of the gas continues to fall inward and is accreted by the black hole, which drives a wind outward into the galaxy. Once the black hole grows large enough, this wind unbinds the remaining gas and halts the star formation episode, at least until another major accretion event occurs. This scenario naturally explains many aspects of the low-redshift Universe (such as the relation of black holes to spheroids rather than to disks), but its application to very high redshifts—where spheroids may or may not even exist, and the much more rapid growth of galaxy-sized halos likely prevents active black holes from entirely halting star formation—is far from clear.

8.8 From Galaxy Model to Stellar Spectra

In addition to the raw star formation rate, most observables depend on the initial mass function (IMF) of the stars. We discussed this in sections 5.2.4 and 5.3.3, where we described how local measurements are consistent with a (broken) power law in the stellar mass range of $\sim$0.1–100 $M_\odot$. Once true galaxies form, with reasonably enriched gas, the IMF likely approaches this form, though (as we argued before) the characteristic mass may increase at higher redshifts owing to the higher CMB temperature.

An additional issue that appears to be important for generating realistic stellar populations is the finite mass of the gas clouds from which stars form, as it now appears that most stars form in groups (though they may later disperse). The range of allowable star cluster masses is called the *cluster initial mass function*; local observations are consistent with a power-law distribution of the cluster number count per unit mass of slope ~ -2 (reflecting an equal amount of mass per logarithmic mass bin) between a few tens of solar masses to $\sim 10^6$ $M_\odot$.[17] These clusters are important because the total fuel mass may limit the maximum stellar mass that can form in that environment—in other words, even if the underlying stellar IMF stretches smoothly to very high masses, a dearth of high-mass gas clouds will cause a dearth of high-mass stars. In the past, the cluster sites were generally ignored, and a population of stars was generated by drawing from the IMF over a uniform range of stellar masses. More accurate stellar population models can be constructed by stochastically sampling the cluster IMF to generate a set of star clusters and then stochastically sampling the stellar IMF (taking into account the maximum stellar mass allowed within each one).[i]

Given a metallicity and following this procedure for a set of stars forming at a particular instant, one can then calculate how the luminosity and spectrum

[i] To many readers, it might seem more natural to define the IMF as the net result of this process, since that would provide the galaxy-wide IMF of stars relevant for cosmologists. However, to measure the stellar IMF, one must find a population of stars that formed simultaneously—in other words, a single cluster. Thus, the canonical stellar IMF—measured long before the importance of the cluster IMF was recognized—is only part of the "real" mass distribution of stars.

of the population evolve with time using libraries of stellar models. Although isolated nonrotating stars are well understood, there remain some important uncertainties in this modeling. The fundamental challenge is that the ionizing luminosity comes from only a small fraction of the stars (those with the highest masses). Thus, small variations in their formation efficiency or properties can cause substantial uncertainties in the models.

For example, the ionizing flux of stars cannot be observed directly and instead must be modeled from their feedback effects on surrounding H II regions. Meanwhile, the massive stars responsible for these photons have atmospheres that are out of local thermodynamic equilibrium and often undergo substantial mass loss through line-driven winds. These so-called Wolf-Rayet stars present particular challenges to models.

As a second example, most (>75%) stars are born with neighbors (as binaries or even larger multiple systems) in the local Universe. Binarity can dramatically affect the evolution of the component stars. For example, suppose the more massive star reaches its supergiant phase first. It expands rapidly, with some of its envelope passing the Roche limit and escaping. The "naked" surface of this more massive star then becomes hotter, producing more ionizing photons. Meanwhile, the neighbor may accrete some of its neighbor's lost mass and itself become more massive (and hence hotter) and possibly gain angular momentum and rotate faster (which also tends to make it hotter).

Overall, different stellar models vary by a factor of a few in their ionizing flux, even at a fixed metallicity and stellar IMF.[18] They are generally more consistent with one another at longer wavelengths, but the differences can still be important. Nevertheless, the following general trends are apparent:

Awkward spacing / PE

- ***Stellar age:*** Because the most massive stars have the shortest lifetime, the spectrum (particularly at high frequencies) is extremely sensitive to the elapsed time since a star formation episode. Figure 8.6(a) shows this explicitly. After only 1 Myr, many stars have not yet evolved into their hot phase, and so the ionizing flux is relatively small. The ionizing spectra harden shortly afterward and then rapidly fade away as the hot stars die. Meanwhile, the continua also fade steadily as more stars fade into white dwarves or explode in supernova.

Awkward spacing / PE

- ***Star formation history:*** A corollary of the previous point is that spectral measurements can determine the star formation history of a galaxy. There is, however, the important possibility that star formation may not be instantaneous. If it instead continues at a constant rate for a long time period (i.e., much longer than the age of the most massive stars), the high-energy photons will still be sensitive to the high-mass, short-lived stars (and hence the current star formation rate), but the lower-energy photons will depend on the integrated population of low-mass stars and so will measure the total stellar mass. Figure 8.6(b) shows spectra with ongoing star formation over timescales of 10^6, 10^7, 10^8, 10^9, and 10^{10} years (from lower to upper curves). Note how the spectra roughly converge after long times, only

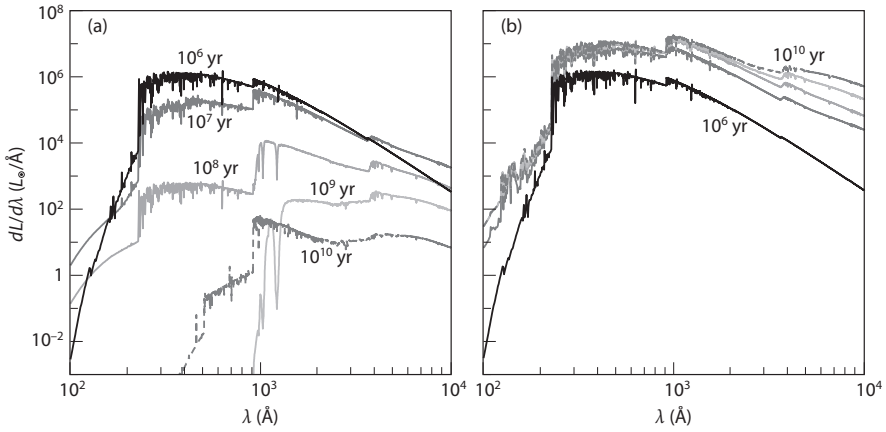


Figure 8.6 Spectral synthesis models of stellar populations. *(a)* Spectra for an instantaneous burst of star formation with $M_\star = 10^6\ M_\odot$. *(b)* Spectra for a constant star formation rate with $\dot{M}_\star = 1\ M_\odot\ \mathrm{yr}^{-1}$. Both panels show predicted spectra for populations 10^6, 10^7 10^8, 10^9, and 10^{10} yr after the onset of star formation. The calculation assumes $Z = 0.05\ Z_\odot$, includes binaries, and ignores nebular reprocessing in all cases. It adopts an IMF with a slope of -1.3 for 0.1–0.5 $M_\odot$, and -2.35 for 0.5–120 $M_\odot$, and does not account for the finite mass of star-forming clouds. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

increasing at very long wavelengths as the galaxy continues to accumulate more and more low-mass stars.

However, we should emphasize that the starburst and constant star formation rate histories are only simple examples; more detailed observations can constrain more complex histories. For the high redshifts of interest to us, where galaxies grow extremely rapidly, so-called exponential histories, where SFR $\propto e^{t/t_\star}$ when smoothed over cosmological times, may also be appropriate. However, any single accretion event may lead to a burst of star formation that dies off rapidly with time, so that SFR $\propto e^{-t/t'_\star}$.

Awkward spacing / PE

- ***Metallicity:*** In general, the higher opacities of heavy elements lead to slightly cooler stellar atmospheres and hence redder spectra. Of course, they also change the spectral lines substantially. Figure 8.7 illustrates this relationship for low- and high-metallicity models (see also §5.4 for a comparison with Population III models). Although the long-wavelength tail is nearly unchanged, increasing the metallicity decreases the ionizing flux by a substantial factor. The nontrivial differences among these spectra indicate that the metallicity is an observable quantity given high-resolution spectra. However, one must bear in mind that metallicity is likely to evolve as star formation proceeds, since it is the stars themselves that enrich the medium, so the cumulative stellar population in a galaxy may not be well described by a single metallicity.

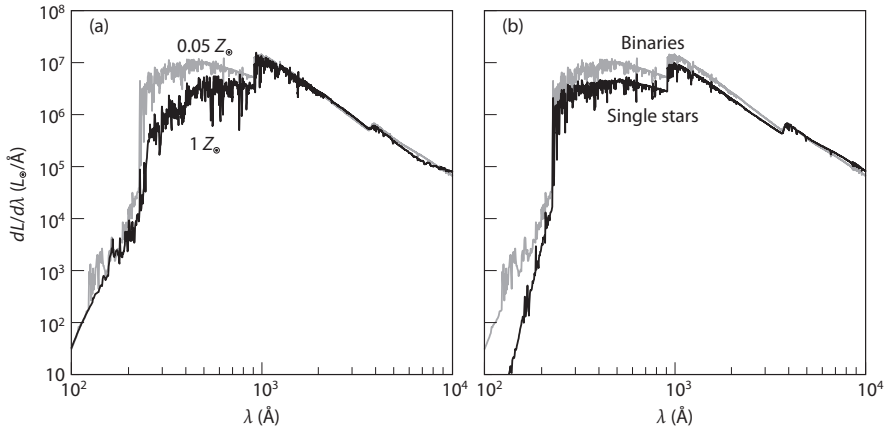


Figure 8.7 Spectral synthesis models of stellar populations: *(a)* variation with metallicity; *(b)* contribution of stellar binaries (assuming $Z = 0.05\ Z_\odot$). Both panels show predicted spectra for a constant star formation rate of 1 $M_\odot\ \mathrm{yr}^{-1}$, 10^8 yr after the star formation began. The calculation ignores nebular reprocessing in all cases, adopts an IMF with a slope of -1.3 for the stellar mass range 0.1–0.5 $M_\odot$ and a slope of -2.35 for masses between 0.5 and 120 $M_\odot$. It does not account for the finite mass of star-forming clouds. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

Awkward spacing / PE

- ***Binaries:*** Finally, we have already mentioned that the inclusion of binary evolution can substantially modify the far-UV fluxes of stellar populations. Figure 8.7(b) shows this effect explicitly. Binaries only slightly change the long-wavelength flux but increase the ionizing flux significantly.

8.9 Signatures of the Interstellar Medium

In the previous section we saw how "synthesized" galaxy spectra can be created given information about the stars and their formation history. Of course, the other major component of the galaxy—its ISM—also has important observable consequences that can affect both the observed continuum of the stars and, especially, the galaxy's spectral lines. A firm grasp of these effects is necessary to understand the stellar component, but it also allows us to learn about the diffuse component of the galaxy and hence its fuel supply and feedback processes. Here, we briefly outline the most important of these effects.

8.9.1 Nebular Emission Lines

The raw stellar spectra computed in §8.8 likely do not reach an observer without substantial modification by their surroundings. The most immediate is the interaction of ionizing photons with the local ISM: presuming that the stars

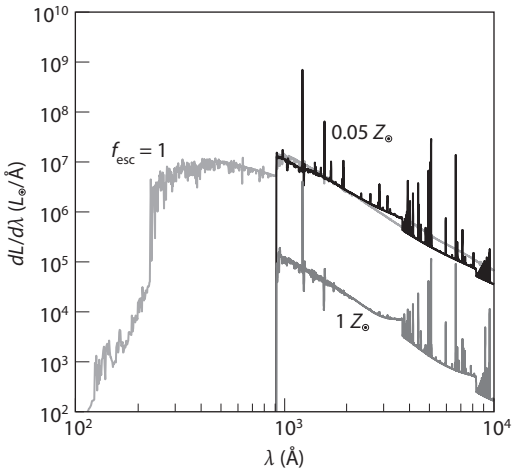


Figure 8.8 Effects of host galaxy absorption on stellar spectra. The curve extending to short wavelengths is the raw stellar spectrum (identical to the curve with binaries in Figure 8.7(b). The other two curves show the spectrum assuming that all ionizing photons are absorbed by the galaxy ISM and reprocessed into emission lines at longer wavelengths. The upper curve assumes $Z = 0.05\ Z_\odot$; the lower curve assumes solar metallicity for both the stars and ISM. The latter is shifted down by two orders of magnitude for clarity of presentation. Generated using the BPASS population synthesis code (http://www.bpass.org.uk).

form in dense environments, many of those photons will be absorbed by local hydrogen or helium atoms. We typically parameterize the fraction that escape their host galaxy as f_{esc}, which is at most a few percent in low redshift galaxies (see §9.2.2). The remaining photons ionize atoms in their host galaxy, which then undergo radiative cascades, reprocessing the energy originally contained in ionizing photons into emission lines at longer wavelengths.

Figure 8.8 shows two examples of this reprocessing, for two different assumed metallicities (note that the solar metallicity curve has been shifted down by a factor of 100 for clarity of presentation; its continuum amplitude is in reality just slightly smaller than that for the low-metallicity model). The strengths of these recombination lines are determined by ionization balance in the H II regions. Assuming that the regions are Strömgren spheres (see §9.1), the total number of recombinations per second is equal to the total number of ionizations, so the emission lines measure the ionizing luminosity. The relative strengths of the hydrogen lines (and helium lines, for very low metallicities and hot stars) depend on atomic physics and so provide a measure of the temperature of the gas.

However, metal lines can also be important diagnostics, if they exist. These are usually collisionally excited forbidden transitions, such as [O II], [O III], and [N II]; they are important because such lines have excitation temperatures

$\sim 10^4$ K, comparable to the expected temperatures of stellar H II regions. The ratios of the strength of these emission lines to those of hydrogen depend on the (gas) metallicity and can be used to estimate it; this dependence has proved to be very useful at lower redshifts, though it is not yet possible at $z > 4$.

Nebular emission lines also offer a useful probe of the escape fraction, because their strength is proportional to $(1 - f_{\rm esc})$. Reprocessing shifts photons from the short-wavelength tail to longer wavelengths and so can even change broadband colors (i.e., a significant fraction of the energy measured in a particular observational filter may be contained in emission lines rather than in the raw stellar continuum). For example, suppose one estimates a spectral index $f_\lambda \propto \lambda^{-\beta}$ from the average broadband colors. The difference between full nebular reprocessing and full escape corresponds to a range in $\beta \sim 2.2$–3.1 for very young stellar populations (<3 Myr), though the difference falls to ~ 0.1 for older populations (>100 Myr).[19] This difference is particularly useful, because even when $f_{\rm esc} \sim 1$, photons with $\lambda < 912$ Å will not be directly observable, owing to absorption by the intervening IGM at $z > 5$. The longer-wavelength emission lines may thus provide the best direct diagnostics of this important quantity.

8.9.2 Dust

The other obvious effect on the stellar spectrum comes from dust, which absorbs stellar photons (especially those with short wavelengths), heats up, and ultimately reradiates that energy in the IR or submillimeter bands. The effects of dust depend on its total mass, its composition, and the relative geometry of the stellar and dust components of the ISM. The total dust mass determines the overall extinction of the gas, while its composition determines the relative extinction across different wavelengths. Unfortunately, this extinction law is found to vary even among nearby galaxies, particularly for short wavelengths. Given that high-redshift galaxies exist much earlier in their life cycle than nearby objects, one would also expect their dust to have very different compositions from those in the present-day Universe. Moreover, if the dust preferentially surrounds star-forming regions, it will have a larger effect on young, hot stars than on the low-mass stars that may have wandered far from their birth sites. Thus, predicting the dust absorption from early galaxies is rather difficult.

The dust emission is equally interesting. Each particle will radiate thermally, though the spectrum will not typically be a true blackbody because dust in different environments may vary in temperature (in the Milky Way, ranging from ~20 to 40 K in the low-density ISM up to several hundred Kelvins in star-forming regions) and because dust particles range in size (the blackbody approximation is not valid for wavelengths smaller than the particle radius). In a simple model, the dust emission spectrum can be parameterized by two quantities: *(i)* the dust temperature T_d and *(ii)* the dust emissivity $\epsilon_{\nu,\rm dust}$.

The dust temperature T_d is set by balancing the incident energy against the dust emission. In the simplest model, we assume blackbody emission and

write

$$T_d^4 \approx T_{\rm CMB}^4 + T_m^4 + T_\star^4 + T_{\rm AGN}^4, \tag{8.15}$$

where the four terms account for the CMB radiation field, the nonradiative energy input (via cosmic rays or supernovae), the stellar radiation field, and any energy input from AGN (which appears to be important in some galaxies). The last two quantities presumably scale with the surface density of star formation and black hole accretion rate, respectively, though locally they appear to saturate at ∼60 K and ∼150 K, respectively.[20] The CMB contribution is rarely important at the present day, but becomes much more significant at higher redshifts.

The dust emissivity is often approximated as a power law, $\epsilon_{\rm dust}(\nu) \propto \nu^\beta$, with $\beta \sim 1$ at high photon frequencies ν (to match observations) and $\beta \to 2$ at long wavelengths from standard scattering theory. If the dust is optically thin, the spectrum will follow $f_{\nu,\rm dust} \propto \epsilon_{\nu,\rm dust} B_\nu$, and the normalization will be determined by balancing the input luminosity (from stars or AGN) with this thermal emission. At low and moderate redshifts, some very rapidly star forming galaxies have such high dust content that nearly all their radiative output emerges in the IR and submillimeter bands. Whether more distant analogs for these exist is so far unknown and depends primarily on how quickly galaxies can build large dust columns.

Although it is clearly difficult to predict from first principles, this dust emission has one very important property from an observer's perspective: the spectra of dusty star-forming galaxies are such that in the submillimeter band, the observed fluxes will be nearly independent of redshift well into the cosmic dawn. This phenomenon occurs because the peak of the blackbody spectrum usually lies blueward of the observational bands, so it moves into the observed bands as the galaxy's redshift increases. Such a *negative K-correction* makes submillimeter observations potentially extremely powerful for observing distant galaxies. Figure 8.9 illustrates this for a model galaxy based on a local composite of dust-dominated galaxies. It shows how the observed flux for galaxies in three different bands varies with redshift for three fiducial dust temperatures. Interestingly, at the longest wavelengths and/or lowest dust temperatures, the flux hardly varies with redshift. If a telescope (such as ALMA) can detect a given galaxy population at $z \sim 1$, it may be able to detect a similarly dusty galaxy the way to $z \sim 10$.

8.9.3 Interstellar Absorption Lines

In addition to metal emission lines from H II regions (see §8.9.1), a galaxy spectrum will also have absorption lines owing to heavy elements in the ISM. In principle, these are interesting for measuring the gas-phase metallicity of the ISM; however, in lower-redshift galaxies the strongest lines tend to be saturated (as can be noted through the relative strengths of doublet lines), which makes such a measurement extremely difficult.[21]

Instead, these absorption lines are useful for measuring the properties of galactic winds. Interestingly, although many of the lines appear saturated, they

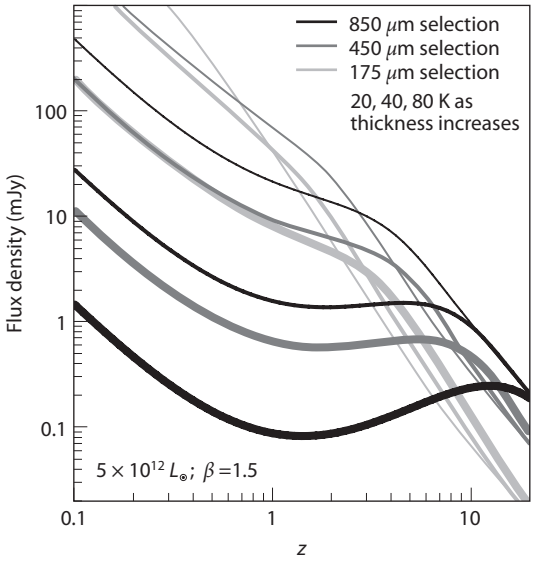


Figure 8.9 Observed flux density as a function of redshift in three submillimeter bands, for several different fiducial dust temperatures. The solid, dotted, and dashed lines assume observations in the 850, 450, and 175 μm bands, respectively. The three curves within each set take different dust temperatures, $T = 20$, 40, and 80 K, from thick to thin lines. All assume that the dust has a power-law emissivity with index $\beta = 1.5$. Blain, A. W., et al., *Phys. Rep.* **369**, 111 (2002). Copyright 2002 by Elsevier.

do *not* completely attenuate the starlight. The depth of the absorption therefore tells us the covering fraction of the high-metal-column gas. Meanwhile, these absorption lines are nearly always redshifted, as would be expected for gas flowing out of the galaxy along the line of sight toward the observer. These lines (together with Lyman-α, which we discuss in chapter 11) provide the best direct evidence for galactic outflows at lower redshifts. However, their interpretation remains extremely difficult because they provide no information about the distance the gas has traveled from the galaxy.

8.9.4 Radio Emission Lines

Another important tracer of the gas phase is emission from molecular and atomic lines: these provide a significant fraction of the cooling radiation that escapes galaxies, especially in star-forming regions. We describe two important examples here: CO, which is an excellent tracer of star formation in the local Universe (and at moderate redshifts), and the [C II] fine-structure line (with a rest wavelength of 157.7 μm), which contains $\sim$0.1%–1% of the bolometric luminosity of nearby star-forming galaxies. Table 8.1 lists many other possible transitions, together with their approximate (local) relation between luminosity and star formation rate.

Eliminate leading space

Table 8.1 Prominent Interstellar Emission Lines in Star-Forming Galaxies, along with Their Typical Ratio R between the Luminosity and Star Formation Rate [in units of $L_\odot(M_\odot \mathrm{yr}^{-1})^{-1}$]. For the first seven lines R is measured from a sample of low redshift galaxies; the other lines have been calibrated based on the galaxy M82. E. Visbal, & A. Loeb, *JCAP* **11**, 16 (2010).

Species	Emission wavelength (μm)	$R[L_\odot\ (M_\odot\ \mathrm{yr}^{-1})^{-1}]$
C II	158	6.0×10^6
O I	145	3.3×10^5
N II	122	7.9×10^5
O III	88	2.3×10^6
O I	63	3.8×10^6
N III	57	2.4×10^6
O III	52	3.0×10^6
^{12}CO(1–0)	2610	3.7×10^3
^{12}CO(2–1)	1300	2.8×10^4
^{12}CO(3–2)	866	7.0×10^4
^{12}CO(4–3)	651	9.7×10^4
^{12}CO(5–4)	521	9.6×10^4
^{12}CO(6–5)	434	9.5×10^4
^{12}CO(7–6)	372	8.9×10^4
^{12}CO(8–7)	325	7.7×10^4
^{12}CO(9–8)	289	6.9×10^4
^{12}CO(10–9)	260	5.3×10^4
^{12}CO(11–10)	237	3.8×10^4
^{12}CO(12–11)	217	2.6×10^4
^{12}CO(13–12)	200	1.4×10^4
C I	610	1.4×10^4
C I	371	4.8×10^4
N II	205	2.5×10^5
^{13}CO(5–4)	544	3900
^{13}CO(7–6)	389	3200
^{13}CO(8–7)	340	2700
HCN(6–5)	564	2100

CO has a ladder of rotational levels $J \to (J-1)$ with frequencies $\nu_J = J\nu_{\rm CO}$, where $\nu_{\rm CO} = 115.3$ GHz, which corresponds to an excitation temperature of $T_{\rm CO} = 5.5$ K. This low temperature means that CO is excited even in the cold, dense molecular clouds out of which stars form. Moreover, because carbon and oxygen are relatively abundant, it is by far the strongest metal line in such regions. At low to moderate redshifts, there is a tight correlation between CO luminosity (here expressed in the 1–0 transition) and the star formation rate,[22]

$$L_{\rm CO(1-0)} = 3.2 \times 10^4 L_\odot \left(\frac{\rm SFR}{M_\odot\ \mathrm{yr}^{-1}}\right)^{3/5}. \tag{8.16}$$

As usual, it is not clear whether this relation can safely be extrapolated to high redshifts. In fact, the astute reader may notice that equation (8.16) is inconsistent with our assumption that $L_{CO} \propto$ SFR in Table 8.1: different galaxy samples and different conversions from luminosity to SFR yield different results, and hence the scaling to higher-redshift galaxies is especially difficult.

To predict the CO luminosity on more physically motivated grounds, we need to know the molecule's abundance as well as its excitation temperature. The latter is set by the cloud's dust (T_d in equation 8.15). The metallicity *may* not be as important as it seems: in local galaxies, giant molecular clouds are optically thick in CO, so decreasing the CO content does not (at first) decrease the overall luminosity. However, ~~note that the dust temperature in equation (8.15) does implicitly depend on the metal content, because once a cloud becomes optically thin to the stellar photons, the dust temperature decreases. This in turn decreases~~ the CO luminosity. In fact, nearby low-metallicity galaxies fall well below the relation in equation (8.16), though the much more compact high-redshift galaxies may have very different characteristics.[23]

Moreover, because T_{CO} is so small, many individual levels can be excited and many transitions can be visible. In local thermodynamic equilibrium, the line ratios depend only on temperature, but at different temperatures and densities the higher levels may not be thermalized. Ideally, one would then like to observe a wide range of lines to fully characterize the molecular clouds.

An alternative bright probe is the fine-structure 157.7 μm line of [C II], which is much less sensitive to the chemistry of the molecular clouds. This line, which arises from a ${}^2P_{3/2} \rightarrow {}^2P_{1/2}$ electronic transition, has an excitation temperature set primarily by collisions with free electrons and interactions with CMB photons, so it can be predicted much more robustly; the primary uncertainty is simply the mass of atomic carbon, or the metallicity of the gas. For $z > 6$, the [C II] line is redshifted into the submillimeter or millimeter range and may be observed with the ALMA telescope.

We will return to radio line emission in §13.2.2, where we ~~discus~~ its utility in comparison with that of other probes.

discuss / AA

8.10 Gravitational Lensing

We end this chapter by describing the physics behind an important technique for understanding the matter content of galaxies, one that is provided for free by nature: *gravitational lensing*. 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 held by a strongly lensed galaxy at $z = 7.6$.[24] As of the writing of this book, this method has provided candidate galaxies with possible redshifts up to $z \sim 10$, though without further spectroscopic confirmation to make these detections robust. We will return to the utility of these lenses in chapter 10; for now we focus on the theoretical background.

if the dust opacity decreases far enough, the interstellar UV background can dissociate the CO, decreasing / AA

The chance alignment of a foreground object along the line of sight to a high redshift source could result in the magnification, distortion, and potential splitting of the source image owing to the deflection of its light rays by the gravitational field of the foreground object. The probability for gravitational lensing grows with increasing source redshift, because of the increase in the path length of the source photons. Although the lensing probability (<1%) is only of anecdotal significance for sources at $z < 2$, its magnitude could rise by an order of magnitude and affect the statistics of bright sources during the epoch of reionization.

Assuming that the gravitational potential of the lens is nonrelativistic, $|\Phi/c^2| \ll 1$, the effect of space–time curvature on the propagation of light rays is equivalent to a medium with an effective index of refraction n,

$$n = 1 - \frac{2}{c^2}\,\Phi. \tag{8.17}$$

This result follows from the deviation imparted to the phase of the electromagnetic wave by the potential of the lens (relative to a flat space–time metric). The lens potential Φ is negative and approaches zero at infinity. As in normal geometric optics, a refractive index $n > 1$ implies that light travels slower than in vacuum. Thus, the effective speed of a ray of light in a gravitational field is

$$v = \frac{c}{n} \simeq c - \frac{2}{c}\,|\Phi|. \tag{8.18}$$

The total time delay Δt, the so-called Shapiro delay, is obtained by integrating over the light path from the observer to the source:

$$\Delta t = \int_{\text{source}}^{\text{observer}} \frac{2}{c^3}\,|\Phi|\,dl. \tag{8.19}$$

A light ray is defined as the normal to the phase front. Since Φ and hence the phase delay of the electromagnetic wave varies across the lens, a light ray will be deflected by the lens as in a prism. The deflection is the integral along the light path of the gradient of n perpendicular to the light path, that is,

$$\hat{\boldsymbol{\alpha}} = -\int \nabla_\perp n\,dl = \frac{2}{c^2}\int \nabla_\perp \Phi\,dl. \tag{8.20}$$

Note that $\hat{\boldsymbol{\alpha}}$ is not a unit vector; rather, the hat is conveniently used to differentiate it from the reduced deflection angle defined in equation (8.28). In all cases of interest the deflection angle is very small. We can therefore simplify the computation of the deflection angle considerably if we integrate $\nabla_\perp n$ not along the deflected ray but along an unperturbed light ray with the same impact parameter (with multiple lenses, one takes the unperturbed ray from the source as the reference trajectory for calculating the deflection by the first lens, the deflected ray from the first lens as the reference unperturbed ray for calculating the deflection by the second lens, and so on).

The simplest lens is a point mass M with a Newtonian potential,

$$\Phi(b, z) = -\frac{GM}{(b^2 + z^2)^{1/2}}, \tag{8.21}$$

where b is the impact parameter of the unperturbed light ray, and z indicates distance along the unperturbed light ray from the point of closest approach. We therefore have

$$\nabla_\perp \Phi(b, z) = \frac{GM\mathbf{b}}{(b^2 + z^2)^{3/2}}, \tag{8.22}$$

where $\mathbf{b}$ is orthogonal to the unperturbed ray and points toward the point mass. Equation (8.22) then yields the deflection angle

$$\hat{\boldsymbol{\alpha}} = \frac{2}{c^2} \int \nabla_\perp \Phi \, dz = \frac{4GM}{c^2 b}. \tag{8.23}$$

Since the Schwarzschild radius is $R_{\rm Sch} = (2GM/c^2)$, the deflection angle is simply twice the inverse of the impact parameter in units of the Schwarzschild radius. As an example, the Schwarzschild radius of the Sun is 2.95 km, and the solar radius is 6.96×10^5 km. A light ray grazing the limb of the Sun is therefore deflected by an angle of 8.4×10^{-6} rad $= 1.7$ arcsec.

lowercase italics r / AA

The deflection angle from a more complicated mass distribution can be treated as the sum of the deflections caused by the infinitesimal point mass elements that make the lens. Since the deflection occurs on a scale $\sim b$, which is typically much shorter than the distances between the observer and the lens or the lens and the source, the lens can be regarded as thin. The mass distribution of the lens can then be replaced by a mass sheet orthogonal to the line of sight, with a surface mass density

$$\Sigma(\boldsymbol{\xi}) = \int \rho(\boldsymbol{\xi}, z) \, dz, \tag{8.24}$$

where $\boldsymbol{\xi}$ is a two-dimensional vector in the lens plane. The deflection angle at position $\boldsymbol{\xi}$ is the sum of the deflections from all the mass elements in the plane:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{4G}{c^2} \int \frac{(\boldsymbol{\xi} - \boldsymbol{\xi}')\Sigma(\boldsymbol{\xi}')}{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2} \, d^2\xi'. \tag{8.25}$$

In general, the deflection angle is a two-component vector. In the special case of a circularly symmetric lens, the deflection angle points toward the center of symmetry and has an amplitude

$$|\hat{\boldsymbol{\alpha}}(\xi)| = \frac{4GM(\xi)}{c^2 \xi}, \tag{8.26}$$

where ξ is the distance from the lens center, and $M(\xi)$ is the mass enclosed within radius ξ,

$$M(\xi) = 2\pi \int_0^\xi \Sigma(\xi')\xi' \, d\xi'. \tag{8.27}$$

The basic lensing geometry is illustrated in Figure 8.10. A light ray from a source S is deflected by the angle $\hat{\boldsymbol{\alpha}}$ at the lens and reaches an observer O. The angle between some arbitrarily chosen axis and the true source position is $\boldsymbol{\beta}$, and the angle between the same axis and the image I is $\boldsymbol{\theta}$. The angular diameter

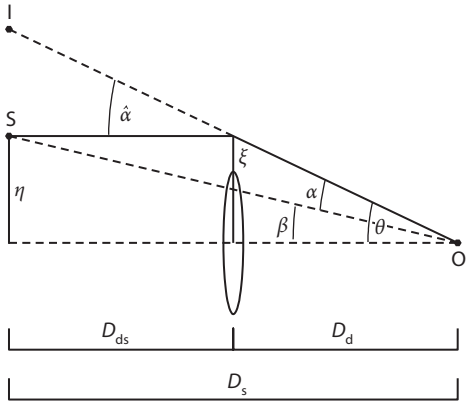


Figure 8.10 Geometry of gravitational lensing. The light ray propagates from the source S at transverse distance η from an arbitrary axis to the observer O, passing the lens at transverse distance ξ. It is deflected by an angle $\hat{\alpha}$. The angular separations of the source and the image from the axis as seen by the observer are β and θ, respectively. The distances between the observer and the source, the observer and the lens, and the lens and the source are D_s, D_d, and D_{ds}, respectively.

distances between observer and lens, lens and source, and observer and source are denoted here as D_d, D_{ds}, and D_s, respectively.

It is convenient to introduce the reduced deflection angle

$$\boldsymbol{\alpha} = \frac{D_{ds}}{D_s}\,\hat{\alpha}. \tag{8.28}$$

The triangular geometry in Figure 8.10 implies that $\theta D_s = \beta D_s - \hat{\alpha} D_{ds}$, so that the positions of the source and the image are related through the simple *lens equation*,

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta}). \tag{8.29}$$

The nonlinear lens equation allows for multiple images $\boldsymbol{\theta}$ at a fixed source position $\boldsymbol{\beta}$. In a flat Universe, the comoving angular-size distances simply add, with $D_{ds}(1+z_s) = D_s(1+z_s) - D_d(1+z_d)$.

Because of the equivalence principle, the gravitational deflection is independent of photon wavelength. In addition, since the phase-space density of photons must be conserved (Liouville's theorem), gravitational lensing preserves the surface brightness of the source and changes only its apparent surface area. The total flux received from a gravitationally lensed image of a source is therefore changed in proportion to the ratio between the solid angles of the image and the source. For a circularly symmetric lens, the magnification factor μ is given by

$$\mu = \frac{\theta}{\beta}\,\frac{d\theta}{d\beta}. \tag{8.30}$$

An extended source is lensed as a sum over infinitesimal (pointlike) segments, each centered on different sky coordinates and having its own magnification factor.

8.10.1 Special Examples of Lenses

8.10.1.1 Constant Surface Density

For a mass sheet with a constant surface density Σ, equation (8.26) implies a reduced deflection angle of

$$\alpha(\theta) = \frac{D_{ds}}{D_s}\frac{4G}{c^2\xi}(\Sigma\pi\xi^2) = \frac{4\pi G\,\Sigma}{c^2}\frac{D_d D_{ds}}{D_s}\theta, \tag{8.31}$$

where $\xi = D_d\theta$. In this special case, the lens equation is linear with $\beta \propto \theta$. Let us define a critical surface-mass density

$$\Sigma_{crit} \equiv \frac{c^2}{4\pi G}\frac{D_s}{D_d D_{ds}} = 0.35\,\mathrm{g\,cm^{-2}}\left(\frac{D}{1\,\mathrm{Gpc}}\right)^{-1}, \tag{8.32}$$

where the effective distance D is defined through the following combination of distances:

$$D \equiv \frac{D_d D_{ds}}{D_s}. \tag{8.33}$$

For a lens with $\Sigma = \Sigma_{crit}$, the deflection angle is $\alpha(\theta) = \theta$, and so $\beta = 0$ for all θ. Such a lens focuses perfectly, with a single focal length. For a typical gravitational lens, however, light rays that pass the lens at different impact parameters cross at different distances behind the lens. Usually, lenses with $\Sigma > \Sigma_{crit}$ can produce multiple images of the source.

8.10.1.2 Circularly Symmetric Lenses

For a circularly symmetric lens with an arbitrary mass profile, equations (8.26) and (8.28) give

$$\beta(\theta) = \theta - \frac{D_{ds}}{D_d D_s}\frac{4GM(\theta)}{c^2\,\theta}. \tag{8.34}$$

A source that lies exactly behind the center of symmetry of the lens ($\beta = 0$) is imaged as a ring. Substitution of $\beta = 0$ in equation (8.34) yields the angular radius of the ring,

$$\theta_E = \left[\frac{4GM(\theta_E)}{c^2}\frac{D_{ds}}{D_d D_s}\right]^{1/2}. \tag{8.35}$$

This *Einstein radius* defines the characteristic angular scale of lensed images: when multiple images are produced, the typical angular separation between them is $\sim 2\theta_E$. Also, sources that are closer than $\sim\theta_E$ in projection (relative to the lens center), experience strong lensing in the sense that they are significantly magnified, whereas sources that are located well outside the Einstein ring

Inconsistent spacing / PE

are magnified very little. In many lens models, the Einstein ring also roughly represents the boundary between source positions that are multiply imaged and those that are only singly imaged. Interestingly, by comparing equations (8.32) and (8.35), we see that the mean surface-mass density inside the Einstein radius is just the critical density $\Sigma_{\rm crit}$.

For lensing by a galaxy mass M at a cosmological distance D, the typical Einstein radius is

$$\theta_{\rm E} = (0.''4) \left(\frac{M}{10^{11}\,M_\odot}\right)^{1/2} \left(\frac{D}{5\ {\rm Gpc}}\right)^{-1/2}, \tag{8.36}$$

with a value that is larger by two orders of magnitude for rich galaxy clusters ($M \sim 10^{15} M_\odot$).

8.10.1.3 Point Mass

For a point mass M the lens equation has the form

$$\beta = \theta - \frac{\theta_{\rm E}^2}{\theta}. \tag{8.37}$$

This equation has two solutions,

$$\theta_\pm = \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4\theta_{\rm E}^2}\right). \tag{8.38}$$

Any source is imaged twice by a point mass lens. The two images are on opposite sides of the source, with one image inside the Einstein ring and the other outside. As the source moves away from the lens (i.e., as β increases), one of the images approaches the lens and becomes very faint, while the other image approaches the true position of the source and asymptotically approaches its unlensed flux.

By substituting β from the lens equation (8.37) into equation (8.30), we obtain the magnifications of the two images,

$$\mu_\pm = \left[1 - \left(\frac{\theta_{\rm E}}{\theta_\pm}\right)^4\right]^{-1} = \frac{u^2+2}{2u\sqrt{u^2+4}} \pm \frac{1}{2}, \tag{8.39}$$

where u is the angular separation of the source from the point mass in units of the Einstein angle, $u = \beta\theta_{\rm E}^{-1}$. Since $\theta_- < \theta_{\rm E}$, $\mu_- < 0$, and so the magnification of the image inside the Einstein ring is negative, implying that the parity of this image is flipped with respect to the source. The net magnification of flux in the two images is obtained by adding the absolute magnifications,

$$\mu = |\mu_+| + |\mu_-| = \frac{u^2+2}{u\sqrt{u^2+4}}\,. \tag{8.40}$$

When the source lies on the Einstein radius, we have $\beta = \theta_{\rm E}$ and $u = 1$, so the total magnification becomes

$$\mu = 1.17 + 0.17 = 1.34. \tag{8.41}$$

8.10.1.4 Singular Isothermal Sphere

A simple model for the mass distribution of a galaxy halo assumes that its collisionless particles (stars and dark matter) possess the same isotropic velocity dispersion everywhere. Surprisingly, this simple model appears to describe extremely well the dynamics of stars and gas in the cores of disk galaxies (whose rotation curve is roughly flat), as well the strong lensing properties of spheroidal galaxies.

We assume a spherically symmetric gravitational potential that confines the collisionless particles that produce it. We can associate an effective "pressure" with the momentum flux of these particles at a mass density ρ,

$$p = \rho\sigma_v^2, \tag{8.42}$$

where σ_v is the one-dimensional velocity dispersion of the particles, assumed to be constant across the galaxy. The equation of hydrostatic equilibrium (which is derived from the second moment of the collisionless Boltzmann equation) gives

$$\frac{1}{\rho}\frac{dp}{dr} = -\frac{GM(r)}{r^2}\ , \quad \frac{dM(r)}{dr} = 4\pi\, r^2\, \rho, \tag{8.43}$$

where $M(r)$ is the mass interior to radius r. A particularly simple solution of equations (8.42) through (8.43) is

$$\rho(r) = \frac{\sigma_v^2}{2\pi G}\,\frac{1}{r^2}. \tag{8.44}$$

This mass distribution is called the *singular isothermal sphere* (SIS). Since $\rho \propto r^{-2}$, the mass $M(r)$ increases $\propto r$, and therefore the rotational velocity of test particles in circular orbits in the gravitational potential is

$$V_c^2(r) = \frac{G\,M(r)}{r} = 2\,\sigma_v^2 = \text{constant}. \tag{8.45}$$

As mentioned previously, this model naturally reproduces the flat rotation curves of disk galaxies.

By projecting the mass distribution along the line of sight, we obtain the surface-mass density

$$\Sigma(\xi) = \frac{\sigma_v^2}{2G}\,\frac{1}{\xi}, \tag{8.46}$$

where ξ is the distance from the center of the two-dimensional profile. The deflection angle from equation (8.26),

$$\hat{\alpha} = 4\pi\,\frac{\sigma_v^2}{c^2} = (1.''16)\,\left(\frac{\sigma_v}{200\,\mathrm{km\,s^{-1}}}\right)^2\ , \tag{8.47}$$

is independent of ξ and points toward the center of the lens. The Einstein radius of the SIS follows from equation (8.35):

$$\theta_{\mathrm{E}} = 4\pi\,\frac{\sigma_v^2}{c^2}\,\frac{D_{\mathrm{ds}}}{D_{\mathrm{s}}} = \hat{\alpha}\,\frac{D_{\mathrm{ds}}}{D_{\mathrm{s}}} = \alpha\,(\theta). \tag{8.48}$$

Owing to circular symmetry, the lens equation is one-dimensional. Multiple images are obtained only if the source lies inside the Einstein ring. If $\beta < \theta_E$, the lens equation has the two solutions

$$\theta_\pm = \beta \pm \theta_E \, . \tag{8.49}$$

The images at $\theta_\pm$, the source, and the lens all lie on a straight line. Technically, a third image with zero flux is located at $\theta = 0$; this image acquires a finite flux if the divergent density at the center of the lens is replaced by a core region with a finite density.

The magnifications of the two images follow from equation (8.30):

$$\mu_\pm = \frac{\theta_\pm}{\beta} = 1 \pm \frac{\theta_E}{\beta} = \left(1 \mp \frac{\theta_E}{\theta_\pm}\right)^{-1} . \tag{8.50}$$

If the source lies outside the Einstein ring (i.e., if $\beta > \theta_E$), there is only one image at $\theta = \theta_+ = \beta + \theta_E$. Searches for highly magnified images of faint galaxies at high redshifts are being conducted near the Einstein radius of clusters of galaxies, where the magnification factor peaks.

Chapter Nine

The Reionization of Intergalactic Hydrogen

The CMB indicates that hydrogen atoms formed 400,000 years after the Big Bang, as soon as the cosmological expansion cooled the gas below 3,000 K. However, observations of the CMB as well as 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 its atomic state back to its constituent protons and electrons in a process known as *reionization*. More specifically, the $z \sim 6$ Lyman-α forest shows that the IGM was highly ionized at that time (see §4.7), though there are possible hints from other methods that some large neutral hydrogen regions persisted until near that time. Thus, we may not need to go to much higher redshifts to begin to see the epoch of reionization. Moreover, CMB polarization studies require that the universe could not have fully reionized earlier than an age of 300 million years (see §13.1.1). It is intriguing that the inferred reionization epoch coincided with the appearance of the first galaxies, which inevitably produced ionizing radiation. *How was the primordial gas transformed to an ionized state by the first galaxies within merely hundreds of million of years?*

We begin this chapter by addressing this question using our tools for describing the formation and evolution of galaxies during the cosmic dawn. The course of reionization can be determined by counting photons from all galaxies as a function of time. Both stars and black holes contribute ionizing photons, but the early Universe was dominated by small galaxies that, in the local Universe, had disproportionately small central black holes. In fact, bright quasars are known to be extremely rare at $z > 6$, so we will generally focus on stellar models as a fiducial case.

Because stellar ionizing photons are only slightly more energetic than the 13.6 eV ionization threshold of hydrogen, they are absorbed efficiently once they reach a region with substantial neutral hydrogen. Consequently, reionization, the IGM was nearly a two-phase medium, characterized by highly ionized zones separated from the neutral sea of gas by sharp ionization fronts. While the redshift at which reionization ended constrains only the overall cosmic efficiency for producing ionizing photons, a detailed picture of these ionized bubbles as they formed and grew will teach us a great deal about the population of the first galaxies that produced this cosmic phase transition.

during / PE

DES: 2 lines under head ok?

9.1 Propagation of Ionization Fronts

The simplest reionization problem to consider is how a single, isolated galaxy ionizes its surroundings. The formation of H II regions, or ionized bubbles,

around galaxies is the fundamental process that drives reionization, although in practice these galaxies are isolated only in the very earliest phases of reionization. Our first goal is to model this problem of an isolated, expanding H II region.

Let us consider, for simplicity, a spherical ionized volume V, separated from the surrounding neutral gas by a sharp ionization front. In the absence of recombinations, each hydrogen atom in the IGM will have to be ionized only once, and the ionized proper volume V_p will simply be determined by

$$\bar{n}_{\rm H} V_p = Q_i, \tag{9.1}$$

where $\bar{n}_{\rm H}$ is the mean number density of hydrogen, and Q_i is the total number of ionizing photons produced by the source.

The size of the resulting H II region depends on the halo that produces it. Let us consider a halo of total mass M_h and baryon fraction Ω_b/Ω_m. To derive a rough estimate, we assume that baryons are incorporated into stars with an efficiency $f_\star$ and that the *escape fraction* for the resulting ionizing radiation is $f_{\rm esc}$. This is the fraction of hydrogen ionizing photons that escape their host galaxy without absorption and so are available to ionize intergalactic gas. We also let $N_{\rm ion}$ be the number of ionizing photons per baryon inside stars; this is $\sim 4{,}000$ for Population II stars with a "present-day" IMF (see §8.8). We finally introduce a parameter $A_{\rm He} = 4/(4 - 3Y_p) = 1.22$, where Y_p is the mass fraction of helium, as a correction factor to convert the number of ionizing photons to the number of ionized hydrogen atoms (assuming that helium is singly ionized as well). This is necessary because the first ionization potential of He I is 24.4 eV, which is sufficiently close to the 13.6 eV required for H I so that typical stellar populations ionize both species together.

At least in our simple model, so far as the IGM is concerned all these parameters are completely degenerate and together determine the overall ionizing efficiency, which we call ζ,

$$\zeta = A_{\rm He} f_\star f_{\rm esc} N_{\rm ion}. \tag{9.2}$$

If we neglect recombinations, then we obtain the maximum comoving radius of the region that the halo of mass M_h can ionize,

$$\begin{aligned} r_{\rm max} = \left(\frac{3}{4\pi}\,\frac{Q_i}{\bar{n}^0_{\rm H}}\right)^{1/3} &= \left(\frac{3}{4\pi}\,\frac{\zeta}{\bar{n}^0_{\rm H}}\,\frac{\Omega_b}{\Omega_m}\,\frac{M_h}{m_p}\right)^{1/3} \\ &= 680\left(\frac{\zeta}{40}\,\frac{M_h}{10^8 M_\odot}\right)^{1/3}\ {\rm kpc}, \end{aligned} \tag{9.3}$$

where $\bar{n}^0_{\rm H}$ is the number density of hydrogen. Here we have taken Population II stars with $f_{\rm esc} = 8\%$ and $f_\star = 10\%$ for a fiducial estimate.

We may make a similar estimate for the size of the H II region around a quasar. For the typical quasar spectrum, $\sim 10^4$ ionizing photons are produced per baryon incorporated into the black hole, assuming a radiative efficiency of $\sim$6% (see §7.2). The overall efficiency of incorporating baryons into the central black hole is low ($<$0.1% in the local Universe; see §7.5.1), but $f_{\rm esc}$ is likely to be close to unity for powerful quasars. Thus, quasars typically have $\zeta > 10$.

This zeroth-order approximation provides a rough guide to the relevant scales. However, the elevated density of the IGM at high redshift implies that recombinations cannot be ignored, so this simplest method must be improved. Just before World War II, the Danish astronomer Bengt Strömgren analyzed the same problem for hot stars embedded in the ISM.[1] 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*) is reached, inside which recombinations balance ionizations:

$$\alpha_B \bar{n}_{\rm H}^2 V_p = \dot{Q}_i, \tag{9.4}$$

where the recombination rate depends on the square of the density and on the recombination coefficient; here we use the case-B value on the assumption that ionizing photons resulting from recombinations to the ground state contribute to the growth of the Strömgren sphere itself (see §4.3).

To model the detailed evolution of an expanding H II region, including a nonsteady ionizing source, recombinations, and cosmological expansion, we write

$$\bar{n}_{\rm H} u_I = \frac{\dot{\mathcal{Q}}_i}{4\pi r_p^2}, \tag{9.5}$$

where u_I is the peculiar velocity of the ionization front (assumed here to be much smaller than the speed of light), and $\dot{\mathcal{Q}}_i$ is the total number of ionizing photons per second that reach the front; it differs from the production rate of ionizing photons, $\dot{Q}_i$, because some of those photons are lost counteracting recombinations within the region. Noting that $4\pi r_p^2 u_I$ is the rate at which proper volume is added to the H II region, which itself also expands with the Hubble flow, we can write[2]

$$\bar{n}_{\rm H}\left(\frac{dV_p}{dt} - 3HV_p\right) = \dot{Q}_i - \alpha_B \langle n_{\rm H} n_e \rangle V_p. \tag{9.6}$$

In this equation, the mean density $\bar{n}_{\rm H} \propto a^{-3}(t)$, and the angular brackets denote a volume average. Note that the recombination rate scales as the square of the density. Therefore, if the IGM is not uniform but contains high-density clumps separated by modestly underdense voids, then the average recombination time will be shorter than in a uniform medium. This difference is often accounted for by introducing a volume-averaged clumping factor C (which is, in general, time dependent), defined by[i]

$$C = \langle n_e^2 \rangle / \bar{n}_e^2. \tag{9.7}$$

Unfortunately, as we will see in §9.2.1, the clumping factor is rather difficult to estimate robustly.

If the ionized volume is large compared with the typical scale of clumping, so that many clumps are averaged over, then equation (9.6) can be solved by

[i]The recombination rate depends on the number density of electrons and hydrogen nuclei; in using equation (9.7) we are neglecting the small contribution made by partially or fully ionized helium.

specifying C. For the comoving volume V, the resulting equation is

$$\frac{dV}{dt} = \frac{1}{\bar{n}_{\rm H}^0}\dot{Q}_i - \alpha_B \frac{C}{a^3}\bar{n}_{\rm H}^0 V. \tag{9.8}$$

The solution for the total number of ionized atoms, $\mathcal{N}_i$, contained in the H II region around a source that turns on at $t = t_i$ can be obtained through an integrating factor,

$$\mathcal{N}_i(t) = \int_{t_i}^{t} \dot{Q}_i(t')\, e^{F(t',t)}\, dt', \tag{9.9}$$

where

$$F(t',t) = -\alpha_B \bar{n}_{\rm H}^0 \int_{t'}^{t} \frac{C(t'')}{a^3(t'')}\, dt'' \,. \tag{9.10}$$

We can simplify this equation in the high-redshift limit ($z \gg 1$), where the scale factor varies as $a \propto t^{2/3}$, if we make the additional assumption of a constant C. Then, defining $f(t) = a(t)^{-3/2}$, we obtain

$$F(t',t) = -\frac{2}{3}\frac{\alpha_B \bar{n}_{\rm H}^0}{\sqrt{\Omega_m} H_0} C \left[f(t') - f(t)\right] = -0.26 \left(\frac{C}{10}\right)\left[f(t') - f(t)\right]. \tag{9.11}$$

We have written equation (9.9) in terms of the number of ionized atoms rather than the volume to emphasize the limits of this approach. One can easily define an *effective* comoving volume $V = \mathcal{N}_i/\bar{n}_{\rm H}^0$. However, one must be careful in applying equation (9.9), because this effective volume V is the volume that would be filled by the ionized gas if held at the mean density *and fully ionized*. The formalism implicitly confines recombinations to the edge of the "ionized volume," rather than allowing for the gas inside the zone to recombine uniformly. This simple model is nevertheless useful for many purposes, especially for steady sources where recombinations are relatively unimportant. We present a more rigorous model for the ionization fronts, and partial ionization inside the H II region, in §9.8.2.

Figure 9.1 shows some examples of the ionized volume evolution for a particular model of an isolated galaxy; the results are scaled to the maximum IGM mass ionized by the galaxy. The models take $\zeta = 40$, which makes $r_{\rm max} \sim 20 r_{\rm vir}$. They also take three possible clumping factors (from top to bottom, $C = 0$, 1, and 10; see §9.2.1) at $z = 10$ and 15 (solid and dashed curves, respectively). For this source, the ionization rate is assumed to be constant for $t_s = 3 \times 10^6$ yr, the characteristic lifetime of the massive stars that produce ionizing photons, before declining $\propto t^{-4.5}$ as these stars die; this is a reasonable approximation to an instantaneous burst of star formation with a "normal" IMF.

Without recombinations, the ionized bubble reaches its maximal size shortly after this characteristic time and remains there at later times; the result here is independent of redshift. If recombinations are allowed, the ionized volume never quite reaches its maximal value, and the shortfall increases with redshift and clumping factor. Moreover, once recombinations are included, the ionized mass shrinks rapidly once the source dims, as recombinations destroy the ionized gas. (We remind the reader again that this does *not* mean that the front

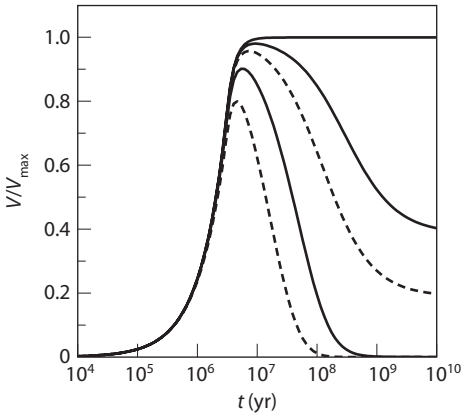


Figure 9.1 Evolution of the effective ionized volume for a stellar ionizing source, scaled to the maximum possible value $V_{max} = 4\pi\bar{n}_H r_{max}^3/3$ (see eq. 9.3). The solid and dashed curves assume that the sources begin shining at $z = 10$ and 15, respectively. Within each set, they take $C = 0$, 1, and 10, from top to bottom. The source has $\zeta = 40$ and is assumed to fade with time $\propto t^{-4.5}$ after a period $t_s = 3\times10^6$ yr, characteristic of the massive star lifetime. Barkana, R., & Loeb, A., *Phys. Rep.* **349**, 125 (2001). Copyright 2001 by Elsevier.

separating ionized and neutral gas shrinks; rather, the recombinations extend throughout the ionized volume, and the front stays more or less in place in this simple model.) Recombinations slow down only at late times as the effective recombination time exceeds the Hubble time.

One additional correction is sometimes necessary for equation (9.6): in the limit of an extremely bright source, characterized by an arbitrarily high production rate of ionizing photons, equation (9.6) would imply that the H II region expands faster than the speed of light. At early times, the ionization front can indeed expand at nearly the speed of light c, but only if the H II region is sufficiently small that the production rate of ionizing photons by the central source exceeds their consumption rate within the current volume. It is straightforward to take the light propagation delay into account. The general equation for the relativistic expansion of the *comoving* radius R of an H II region in an IGM with neutral fraction $x_{H\,I}$ is[3]

$$\frac{dR}{dt} = c(1+z)\left[\frac{\dot{Q}_i - \alpha_B C x_{H\,I}\left(\bar{n}_H^0\right)^2 (1+z)^3\left((4\pi/3)R^3\right)}{\dot{Q}_i + 4\pi R^2 (1+z)\, c x_{H\,I}\bar{n}_H^0}\right], \qquad (9.12)$$

where here $\dot{Q}_i$ is the rate of ionizing photons crossing a shell of the H II region at radius R and time t (and so corresponds to the luminosity of the source at a time in the past). Indeed, for $\dot{Q}_i \to \infty$ the propagation speed of the proper radius of the H II region, $r_p = R/(1+z)$, approaches the speed of light, $(dr_p/dt) \to c$.

9.2 Global Ionization History

The next level of sophistication in understanding reionization is to compute the evolution of the average neutral fraction across the entire Universe. We can obtain a first estimate for the requirements of reionization by demanding one stellar ionizing photon for each hydrogen atom in the Universe. To zeroth order the accounting is relatively simple: the efficiency parameter ζ is simply the number of ionizing photons impacting the IGM produced per baryon inside galaxies; thus, the neutral fraction (ignoring recombinations) is

$$Q_{\rm H\ II} = \zeta f_{\rm coll}, \tag{9.13}$$

where $Q_{\rm H\ II}$ denotes the average *filling factor* of ionized bubbles (i.e., the fraction of the volume of the Universe inside H II regions), and the collapse fraction $f_{\rm coll}$ is the fraction of matter incorporated in galaxies (typically above some minimum mass threshold determined by cooling and/or feedback; see §3.4 and especially Figure 3.12). This equation assumes instantaneous production of photons, that is, that the timescale for the formation and evolution of the massive stars in a galaxy is relatively short compared with the Hubble time at the formation redshift of the galaxy. The primary inputs to such a model are the net ionizing efficiency ζ and $f_{\rm coll}$, which (given a halo mass function) depends on the threshold halo mass that allows star formation. 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 (3.32).

Again, we can improve this simple prescription by accounting for recombinations. To do so, we treat each ionizing source as producing an isolated bubble and assume that the bubble volumes add to give the total filling factor; although, in fact, overlap is very important, this is not a bad approximation because—neglecting internal absorption—any photons that pass into another ionized bubble propagate to its edge and help grow it. Starting with equation (9.8), if we assume a common clumping factor C for all H II regions, 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, we can replace V by the filling factor and $\dot{Q}_i$ by the ionizing photon production rate at time t per unit volume. The latter quantity is simply $\zeta \bar{n}\, df_{\rm coll}/dt$, which provides the emissivity of ionizing photons. Under these assumptions we convert equation (9.8), which describes individual H II regions, to an equation that statistically describes the transition from a neutral Universe to a fully ionized one:[ii]

$$\frac{dQ_{\rm H\ II}}{dt} = \zeta \frac{df_{\rm coll}}{dt} - \alpha(T)\frac{C}{a^3}\bar{n}_{\rm H}^0 Q_{\rm H\ II}, \tag{9.14}$$

[ii]This equation actually suffers from the same small conceptual inconsistency we encountered in Figure 9.1: recombinations do not necessarily decrease the size of the bubbles (and hence $Q_{\rm H\ II}$) but are instead spread throughout the ionized gas. We take a more careful approach to $Q_{\rm H\ II}$ in the following sections of this chapter.

which admits of the solution (in analogy with equation 9.9),

$$Q_{\rm H\ II}(t) = \int_0^t \zeta \frac{df_{\rm coll}}{dt'} e^{F(t',t)}\, dt', \tag{9.15}$$

where $F(t', t)$ is determined by equation (9.11).

Although this equation appears simple, even at this low level of sophistication it hides a number of uncertain parameters. Not only do the elements of ζ each have large uncertainties, but they may also evolve in time; similarly, the clumping factor C depends on the pattern of ionization in the IGM. We next discuss each of these factors in turn.

9.2.1 Recombinations and the Clumping Factor

Before considering ζ, we first discuss some subtleties of the sink term in equation (9.14). First, the recombination coefficient is somewhat uncertain through both the gas temperature (which depends on nonequilibrium processes during reionization; see §4.3.1) and an environmental factor that determines whether case-A or case-B recombination is more appropriate (see §4.3). However, consider the case in which ionizations (and hence recombinations) are distributed uniformly throughout the IGM. Then, case B is appropriate, because each regenerated photon will soon encounter another IGM atom to ionize. But in the highly ionized low-redshift Universe, most recombinations actually take place inside dense, partially neutral Lyman-limit systems (LLSs) because high-energy photons can penetrate inside these high–column density systems (see §4.4.1). However, the ionizing photons produced after recombinations to the ground state usually lie near the Lyman limit (where the mean free path is small), so they are consumed inside the systems. Thus, these photons do not help ionize the IGM, and case A is more appropriate. Which of these regimes is more relevant depends on the details of small-scale clumping and radiative transfer.

Even more problematic is the clumping factor $C(z)$. It may seem at first that this volume-averaged factor can be computed through numerical simulations. But that requires overcoming several difficult problems: *(i)* tracing the gas distribution with sufficient precision to resolve density fluctuations on the smallest scales; *(ii)* correctly tracing the topology of ionized and neutral gas—because the average must be calculated over *only* the ionized gas; and *(iii)* correctly modeling the evolution of gas clumps during the reionization process itself.

The first problem is obvious: even leaving aside the ISM of each galaxy, as well as the overdense environment surrounding each halo (which is implicitly included in $f_{\rm esc}$ in equation 9.2), the Jeans mass in the cold IGM is $<10^5\ M_\odot$. This allows the formation of a well-defined cosmic web, as well as "minihalos," dense gas clouds that virialize but cannot cool or form stars. But, as we shall see, simulations of reionization must span $\sim$100 Mpc boxes to adequately sample the large H II regions, requiring an enormous dynamic range. Thus, even in simulations, clumping is usually accounted for through a "subgrid" model built from semianalytic techniques or bootstrapped from smaller simulations.

The second problem is perhaps more subtle: how do the sources and absorbers relate to each other, and how does ionization affect the small-scale clumping? For example, if low-density gas is ionized first, $C < 1$ throughout most of reionization, because all the dense gas will remain locked up in neutral, self-shielded systems (which cannot, by definition, recombine). However, on large scales the ionizing sources lie inside overdense regions (sheets and filaments), where the recombination rate is higher than average. The relative importance of these two pools of gas changes as reionization progresses, which makes simplified prescriptions for clumping particularly difficult to develop.

Finally, as the gas is ionized, the thermal pressure increases and the clumps evaporate and fade into the IGM. Studying this problem requires simulations of coupled-gas dynamics and radiative transfer, which (although now possible) is difficult and highly dependent on the particular model of reionization (see §9.9 for a detailed discussion). As an additional difficulty, the prereionization gas temperature is uncertain by a factor of 100 or so, making even the initial clumpiness hard to determine.

Thus, while the introduction of the clumping factor is an essential approximation for many analytic models, its evaluation is rather difficult; we describe more physically motivated approaches in §9.5. Nevertheless, a reasonable and concrete estimate is often useful. A recent fit from simulations that ignores the second and third problems but does resolve the proper scales is[4]

$$C(z) = 27.466 \exp(-0.114z + 0.001328z^2). \tag{9.16}$$

As expected, numerical simulations with radiative transfer and heating find a significantly lower effective value of $C(z)$.[5]

9.2.2 The Ionizing Efficiency

We now move on to the source term in equation (9.14), which has two parts: $df_{\rm coll}/dt$ and the ionizing efficiency ζ. The collapse fraction for a given cosmology depends only on $M_{\rm min}$, the mass threshold for galaxy formation. The most common choice for $M_{\rm min}$ corresponds to a virial temperature $T_{\rm vir} = 10^4$ K, the threshold at which hydrogen line cooling becomes efficient for primordial gas (see Fig. 5.1). Above this mass, cooling and fragmentation into stars is relatively straightforward. Other choices are, however, physically plausible in certain regimes. For example, we have seen that H_2 cooling can allow Population III star formation in much smaller halos, while internal feedback within galaxies (like supernova winds) can strongly suppress star formation in halos near the cooling threshold, effectively raising $M_{\rm min}$.

The factor ζ is even more difficult to pin down. A star formation efficiency $f_\star \sim 10\%$ is reasonable for the local Universe, but so little gas has collapsed by $z = 6$ that these local observations do not directly constrain the high-redshift value. Appropriate choices for Population III stars are even more uncertain. To the extent that each halo can form only a single very massive ($m_\star \sim 10^2\ M_\odot$) star that enriches the entire halo ($>10^6\ M_\odot$), $f_\star \sim (\Omega_m/\Omega_b) m_\star/M_h < 10^{-3}$, though larger values are permissible, especially if metal dispersal is inefficient.

The UV escape fraction is small in both nearby galaxies and those at moderate redshifts, with many upper limits $f_{\rm esc} < 5\%$ and only a few positive detections.[6] Interestingly, $f_{\rm esc}$ shows large variance among galaxies; most likely, ionizing photons are able to escape only along clear channels in the galactic ISM, which appear to be quite rare in the objects we can study. However, it could be considerably larger inside small, high-redshift galaxies, whose interstellar media can easily be shredded by radiation pressure, winds, and supernovae, clearing out large escape paths.

$N_{\rm ion}$ depends on the stellar initial mass function and metallicity. Convenient approximations are $N_{\rm ion} \approx 4{,}000$ for $Z = 0.05\ Z_\odot$ Population II stars with a present-day initial mass function, and $N_{\rm ion} < 10^5$ for very massive Population III stars (see §5.4 and 8.8), but these are uncertain by at least a factor of two. In particular, the latter estimate hinges more on the high masses of these stars than on their primordial composition; metal-free stars with a normal Salpeter IMF are only ∼1.6 times more efficient than their Population II counterparts.

Finally, accreting black holes may provide an additional source of ionizing photons that increases the total efficiency of each halo. These sources are likely to have hard spectra and so produce a substantial number of high-energy ionizing photons; thus, they pose particular challenges to understanding reionization, to which we return in §9.8.

Of course, we expect all these factors to evolve throughout reionization due to the feedback processes discussed elsewhere. Thus, a robust model for the filling factor $Q_{\rm H\ II}$ requires a sophisticated understanding of galaxy evolution during the cosmic dawn. This lies well beyond our powers at present, but we can make some progress by generalizing the ionizing efficiency to be a function of both time and halo mass m_h, $\zeta \equiv \zeta(m_h, t)$. The mass dependence is meant to capture internal feedback mechanisms that affect each galaxy in a deterministic fashion, like the effects of starburst winds. With this prescription, we must replace the source term in equation (9.14) with the integral

$$\frac{d}{dt}\int dm_h \frac{m_h}{\bar{\rho}}\zeta(m_h, t)n(m_h, t), \tag{9.17}$$

where $n(m_h, t)$ is the halo mass function (see §3.4). Unfortunately, external feedback mechanisms—which depend on the halo's large-scale environment—require physics inputs and machinery in the model.

additional / PE

9.3 The Phases of Hydrogen Reionization

The process of hydrogen reionization involves several distinct stages (conceptually illustrated in Figure 9.2). The initial "preoverlap" phase 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 that escape from the galaxy itself must then make their way through the surrounding high-density regions, which are characterized by a high recombination rate.

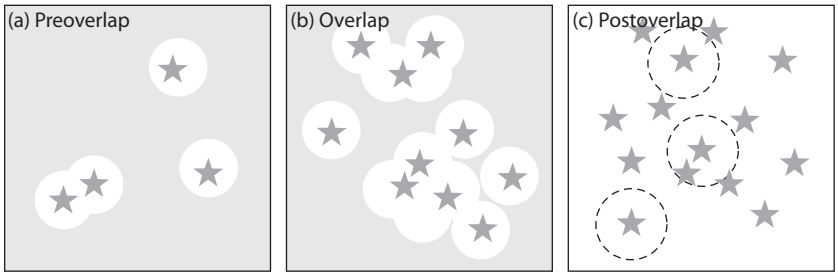


Figure 9.2 Diagram of the three phases of reionization. *(a)* In the *preoverlap* phase, galaxies are uncommon, and their H II regions grow in relative isolation. In practice, galaxies are highly clustered, so even in this stage some of those H II regions meet each other. Nevertheless, the average ionizing background grows fairly slowly during this epoch. *(b)* Once galaxies become sufficiently common, the *overlap* phase, in which ionized bubbles rapidly intersect, begins. During this phase, the ionizing background can increase relatively rapidly as sources are quickly added to each discrete H II region. *(c)* Once nearly all the IGM is ionized, the *postoverlap* phase begins, in which ionizing photons are absorbed by dense clouds of optically thick gas, or Lyman-limit systems. The finite mean free path of ionizing photons, even in a "reionized" universe, is represented here by the dashed circles surrounding a few of the sources.

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 nearly a two-phase medium characterized by highly ionized regions separated from neutral regions by ionization fronts. Furthermore, the ionizing intensity is very inhomogeneous even within the ionized regions.

Because these first sources are highly clustered, this early phase quickly enters the central, relatively rapid "overlap" phase of reionization when neighboring H II regions begin to overlap. Whenever two ionized bubbles join, each point inside their common boundary becomes exposed to ionizing photons from both sources. Therefore, the ionizing intensity inside H II regions rises rapidly during overlap, allowing those regions to expand into high-density gas that had previously recombined fast enough to remain neutral when the ionizing intensity was low. By the end of this stage, most regions in the IGM are able to "see" many individual sources, making the ionizing intensity both larger and more homogeneous than before as the bubbles grow.

During this central phase, most ionizing photons stream through the IGM without absorption, because the gas is highly ionized. However, the protocosmic web makes this gas inhomogeneous, and in dense pockets of the IGM the recombination rate is much larger. These neutral regions—the high-redshift analogs of LLSs (see §4.4.1)—absorb any ionizing photons that strike them, preventing the H II regions from continuing to grow. Eventually, the ionized bubbles become so large that most photons strike one of these LLSs before

reaching the edge of a bubble. This final "postoverlap" phase thus has slower evolution in the ionizing background (at least in the simplest models), modulated by the evaporation of these LLSs, and that background becomes increasingly more uniform.

Of course, this reionization process develops at different rates in different regions of the Universe; naturally, areas with an overabundance of sources undergo more rapid reionization, while those with relatively few sources require input of ionizing photons from external sources. Because the galaxy population traces the underlying density field, these correspond to overdense and underdense regions, respectively. But because the galaxies are highly biased relative to the dark matter, even a modestly overdense region can undergo reionization much earlier. (In fact, if galaxies were unbiased, reionization would not occur any faster in dense regions, because the increased galaxy counts would be exactly canceled by the increased gas density!) Note that this inhomogeneity in the reionization process also means that the three phases identified previously are not clearly distinct from one another: overdense environments rapidly reach the overlap (and even postoverlap) stages while void regions are still in the preoverlap phase.

This general march of reionization from high to low density is referred to as *inside-out* reionization. While most reionization models follow this behavior when averaged over large scales, on sufficiently small scales the process is actually *outside-in*, proceeding from low to high densities, since dense blobs remain partially neutral for a more extended period of time.

Figure 9.3 illustrates this patchiness (or "Swiss cheese topology" as it is often termed). The top left, top center, top right, and bottom left panels show the density of ionized hydrogen (in units of the mean) when $Q_{\rm H\ II} = 25\%$, 50%, 75%, and $\approx$ 100%. The bottom right panel shows the redshift $z_{\rm reion}$ at which each cell in the simulation was ionized. Note the wide distribution of ionized bubble sizes, with the largest bubbles centered around the largest clusters of galaxies in the simulation, and the tight correlation with $z_{\rm reion}$.

9.4 The Morphology of Reionization

Clearly, the patchiness of the ionization field—or its *morphology*—depends sensitively on where galaxies formed at high redshifts. This morphology is therefore of much interest from both theoretical and observational perspectives, and we next describe its theoretical modeling.

Given the complex physics of the sources and sinks of ionizing photons and their interaction in the IGM, it may seem that the problem must be tackled with detailed numerical simulations, and, indeed, much of the early work, beyond the preoverlap stage, followed that approach. However, at its heart, reionization is actually surprisingly straightforward: until the postoverlap stage, it simply requires us to count photons. Thus, a great deal of progress can be made with simple analytic models.

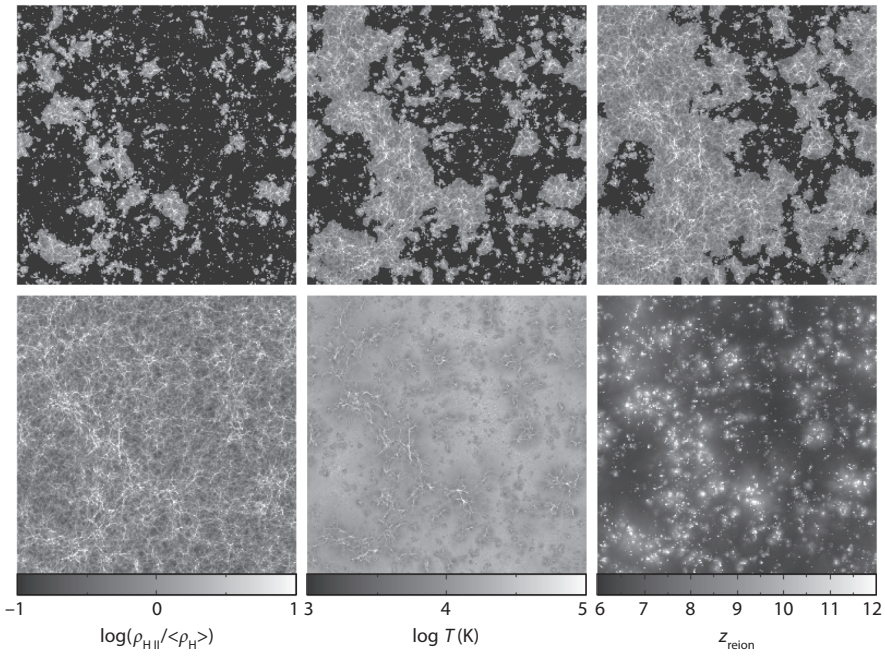


Figure 9.3 Snapshots from a numerical simulation illustrating the spatial structure of cosmic reionization in a slice of 140 Mpc on a side. The simulation describes the dynamics of the dark matter and gas as well as the radiative transfer of ionizing radiation from galaxies. The first four panels (reading across from top left to bottom left) show the evolution of the ionized hydrogen density $\rho_{\rm H\ II}$ normalized by the mean proton density in the IGM $\langle\rho_{\rm H}\rangle$ 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. Trac, H., & Loeb, A. (2010).

Let us consider the simplest possible exercise, shown in the first two panels of Figure 9.4: we count the number of ionizing photons produced by galaxies inside some specified volume of radius R and fractional overdensity δ_R and compare the sum with the number of hydrogen atoms. The region can be ionized only if the former exceeds the latter, or if

$$\zeta f_{\rm coll}(z, \delta_R, R) > 1. \tag{9.18}$$

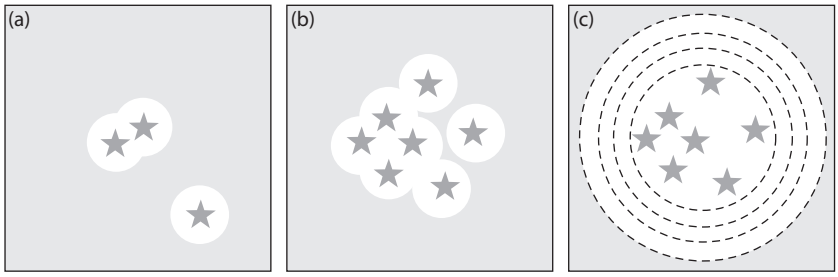


Figure 9.4 Diagram of the excursion set model for generating the ionization morphology. *(a)* In the initial stages, where galaxies are nearly isolated, it is straightforward to compute the ionized volume (white circles) around each source (stars), given only its ionizing efficiency. However, if two such volumes overlap, the overlap region must be distributed over the remaining volume so that ionizing photons are conserved. *(b)* This problem becomes especially acute once a substantial fraction of the IGM is ionized: in that case, redistributing the overlap volume can link up previously separate ionizing bubbles and create a cascading effect. *(c)* The solution is to work from the outside in, rather than trying to begin with each individual galaxy. Here, we compare the number of atoms contained inside the large circle with the number of ionizing photons generated by all the galaxies inside that region. If the latter is larger, the region is ionized; if not, we ask the same question for smaller regions (shown by the dashed circles), until the two quantities balance.

Here $f_{\rm coll}(z, \delta_R, R)$ is the collapse fraction within this region (see equation 3.38 and §3.4.2),

$$f_{\rm coll}(z, \delta_R, R) = {\rm erfc}\left[\frac{\delta_{\rm crit}(z) - \delta_R/D(z)}{\sqrt{2[\sigma^2_{\rm min} - \sigma^2(R)]}}\right], \tag{9.19}$$

where $\delta_{\rm crit}$ is the threshold for halo collapse,[iii] the factor $D(z)$ linearly extrapolates the real fractional overdensity δ_R to the present day for comparison with the collapse threshold, and $\sigma^2_{\rm min}$ is the variance of the density field on the scale corresponding to the minimum mass for galaxy formation, $M_{\rm min}$. The proportionality constant ζ is the ionizing efficiency per baryon in stars (equation 9.2); here we have assumed that it is identical in every galaxy, though that is straightforward to modify, as in equation (9.17).

There are two flaws to this approach. The first is that some fraction of the gas may recombine before the region is completely ionized, so more than one photon per atom is required. If such recombinations were uniform, we could account for them simply by replacing $\zeta \rightarrow \zeta/(1 + N_{\rm rec})$, where $N_{\rm rec}$ is the mean number of recombinations per baryon. In practice, this is not a very good

[iii] Here we have used the constant Press-Schechter criterion for simplicity, but one of the more accurate choices described in §3.4.3 can easily be used.

approximation, so we defer a detailed description of the effects of inhomogeneous recombinations until later.

The second problem is the propagation of photons over large scales. Equation (9.18) is *local*, in that it compares atoms in a region only with photons generated *in the same region*. In fact, a particular patch of space may be entirely ionized by sources from outside the patch. In an extreme example, consider a spherical shell in the IGM that surrounds a galaxy. The galaxy sits inside the shell, but if the shell is sufficiently close to the galaxy, it will nevertheless be ionized.

Thus, to apply equation (9.18), we require some way to adjust the scale R as needed to account for nearby sources. Fortunately, we have already studied just such a technique: the excursion set model for dark matter halos solves this very problem (§3.4.1). In that case, the problem was that a small-scale density fluctuation may lie inside a larger-scale feature that itself may have collapsed to form a halo; in the present case a small region may lie inside a larger ionized bubble. Either way, the solution is to compare the threshold (for spherical collapse or ionization) on *all* scales, working from large to small so as to include neighbors automatically, by phrasing it as a diffusion problem. Figure 9.4(c) illustrates the procedure.

We therefore consider here the trajectory of δ_R as we move from large to small scales. We compare this smoothed overdensity to the criterion in equation (9.18), which we can rewrite as

$$\delta_R > \delta_B(M, z) \equiv \delta_{\rm crit} - \sqrt{2}K(\zeta)[\sigma^2_{\rm min} - \sigma^2(M, z)]^{1/2}, \tag{9.20}$$

where $K(\zeta) = {\rm erf}^{-1}(1 - \zeta^{-1})$, and ${\rm erf}(x) \equiv 1 - {\rm erfc}(x)$. The barrier in equation (9.20) is well approximated by a linear function of σ^2, $\delta_B \approx B(M) = B_0 + B_1\sigma^2(M)$, where B_0 and B_1 are fitting constants. Conveniently, for this linear approximation there is an analytic solution to the diffusion problem (which we derive in detail in §9.4.1), which we can transform into the mass function of ionized bubbles[7]

$$n_b(M) = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M^2}\left|\frac{d\ln\sigma}{d\ln M}\right|\frac{B_0}{\sigma(M)}\exp\left[-\frac{B^2(M)}{2\sigma^2(M)}\right]. \tag{9.21}$$

This function $n_b(M)$ provides the comoving number density of ionized bubbles with IGM mass in the range between M and $M + dM$.

The solid curves in Figure 9.5 show the resulting size distributions for a range of $Q_{\rm H\,II}$ at $z = 15$; the ordinate is the fraction of the ionized volume filled by bubbles of a given size. The most important result of these models is that bubbles grow large during the middle stages of reionization, with characteristic sizes $R_c \sim 1$, 4, 10, and 30 Mpc when $Q_{\rm H\,II} = 0.2$, 0.4, 0.6, and 0.8. Comparing this result with equation (9.3) shows clearly that by the midpoint of reionization a *typical* ionized bubble already contains thousands of sources—overlap is indeed extremely important in determining the morphology of ionized bubbles.

A second important point is the very different shape of these bubble mass functions compared with the halo mass function, which increases toward zero

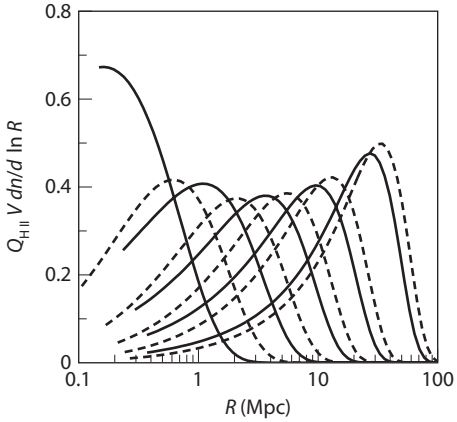


Figure 9.5 H II region size distributions at $z = 15$ in the analytic model of equation (9.21). The solid and dashed curves assume $\zeta \propto m_h^0$ and $m_h^{2/3}$, respectively. From left to right within each set, we take $Q_{\rm H\ II} = 0.05$, 0.2, 0.4, 0.6, and 0.8. Recombinations are assumed to be uniform throughout the IGM. Furlanetto, S. R., McQuinn, M., & Hernquist, L., *Mon. Not. R. Astron. Soc.* **365**, 529 (2006). Copyright 2006 by the Royal Astronomical Society.

mass at all redshifts. The barrier of equation (9.20) increases relatively rapidly toward small M, choking off the formation of small bubbles. This imprints a characteristic size R_c on the ionized bubbles. To understand this size, note that R_c is the scale at which a "typical" density fluctuation is able to ionize itself, without the input of external sources; mathematically, it is where $\sigma(R_c) \approx B$. In the large bubble limit ($B \approx B_0$), our original ionization criterion becomes

$$\zeta f_{\rm coll}(\delta = B_0,\ \sigma^2 = 0) = 1. \tag{9.22}$$

Expanding equation (9.19) to linear order, we can write

$$\sigma(R_c) \approx B_0 \approx \frac{Q_{\rm H\ II}^{-1} - 1}{D(z) b_{\rm eff}}, \tag{9.23}$$

where $b_{\rm eff}$ is the average galaxy bias. Intuitively, a more biased galaxy population provides a larger "boost" to the underlying dark matter fluctuations, allowing larger regions to ionize themselves. The dashed curves in Figure 9.5 illustrate this effect: they show the bubble size distribution if $\zeta \propto m_h^{2/3}$, where m_h is the halo mass. This prescription emphasizes the massive, more biased galaxies and so increases $b_{\rm eff}$. Thus, by measuring the H II region sizes, one can constrain the characteristics of the galaxies driving reionization.

Several properties of equation (9.21) deserve emphasis. First, at a given $Q_{\rm H\ II}$, $n_b(M)$ depends only weakly on redshift. This is because the shape of $f_{\rm coll}(\delta, R)$ evolves only slowly with redshift; quantitatively, $D(z)b_{\rm eff}$ is roughly constant for high-redshift galaxies, assuming that $M_{\rm min}$ is determined by a virial temperature threshold. Second, the width of $n_b(m)$ is ultimately determined by the

shape of the underlying matter power spectrum, which steepens toward larger radii with a shape that is only weakly dependent on astrophysical uncertainties.

Thus, at least in this simple model, the bubble sizes depend essentially on only two parameters: the overall filling fraction of the ionized gas, $Q_{\rm H\ II}$, and the average bias of the ionizing sources, $b_{\rm eff}$. Varying the overall efficiency of reionization (and hence its timing) has only a small effect on the morphology of reionization. This robustness makes the morphology an extremely useful tool for measuring the reionization process.

Finally, the similarity to the Press-Schechter halo mass function also means that most of the machinery used for halo mass functions (for clustering, conditional mass functions, etc.) can be carried over to describe these ionized bubbles. For example, the linear bias of H II regions, defined so that $n_b(m|\delta) = n_b(m)\,[1 + b_{\rm H\ II}(m)\,\delta]$ in a large region of mean overdensity δ, is[iv]

$$b_{\rm H\ II}(m) \approx 1 + \frac{B(m)/\sigma^2(m) - 1/B_0}{D(z)}. \tag{9.24}$$

Note that in this model each bubble must correspond to a region with above-average density (although it can, of course, contain smaller underdense voids). This is obvious from equation (9.18): once the average $f_{\rm coll}(\delta = 0) = 1/\zeta$, the entire Universe must already be ionized.

However, the bias $b_{\rm H\ II}$ can become negative for sufficiently small bubbles. Physically, this occurs because overdense regions are further along in the reionization process, so most small bubbles have already merged with larger H II regions. During the late stages of reionization, only the deepest voids contain galaxies isolated enough to create small bubbles. Nevertheless, the average bias of ionized gas,

$$\bar{b}_{\rm H\ II} \equiv Q_{\rm H\ II}^{-1} \int dm\, n_b(m) V(m) b_{\rm H\ II}(m), \tag{9.25}$$

where $V(m)$ is the volume corresponding to a mass m, is quite large throughout the early stages of reionization, attaining values $\sim$ 3–10.

As another example, each H II region of mass m must have its overdensity equal to the barrier value at $\sigma^2(m)$. One can then generate density trajectories with the initial conditions fixed at these bubblewide values and apply the usual spherical (or ellipsoidal) collapse criterion to generate the conditional halo mass functions within each bubble (cf. eq. 3.43); thus one can predict the galaxy populations that ionize each region of space. We explore this possibility further in §11.7.1.

Finally, we end this section by noting that the *observed* distribution of bubble sizes differs from this "intrinsic" one. The theoretical distribution is evaluated at a single instant in cosmic time; however, real observations observe different

[iv] There is one subtlety in this calculation compared with the usual halo bias. With the linear barrier fit to equation (9.20), the fractional bubble overdensity has a term $B_1\sigma_R^2/B_0$, where σ_R^2 is the mass variance on the large scale on which the bias estimate is made. This term does not scale with the dark matter density and so it spoils a linear bias estimate. Fortunately, it is large only if σ_R^2 is large (i.e., on small scales) or very close to the end of reionization, when $B_1 \gg B_0$.

times because of the finite speed of light.[8] This "light-cone effect" imposes a *maximum* observable bubble size at the end of reionization, which can be estimated via similar arguments to those we have used here. Let us take the slightly simpler case of including only those photons generated within a given region of radius R. Then, the ionization state of that region depends only on the collapsed fraction inside it. Again, reionization should be completed when this fraction exceeds a certain critical value, corresponding to a threshold number of ionizing photons emitted per baryon. There is an offset δz between the redshift at which a region of mean overdensity δ_R achieves this critical collapsed fraction and the redshift $\bar{z}$ at which the Universe achieves the same collapsed fraction on average.

This offset may be computed by expanding equation (9.19) assuming small deviations (an excellent approximation on the large scales and early times relevant here), which gives[9]

$$\frac{\delta z}{(1+\bar{z})} = \frac{\delta_R}{\delta_{\rm crit}(\bar{z})} - \left[1 - \sqrt{1 - \frac{\sigma_R^2}{\sigma_{\rm min}^2}}\right], \tag{9.26}$$

where, again, $\sigma_{\rm min}^2$ is evaluated at the minimum galaxy mass $M_{\rm min}$. Obviously, the offset in the ionization redshift of a region depends on its linear overdensity δ_R. Note also that equation (9.26) is independent of the critical value of the collapsed fraction required for reionization: the only redshift dependence is in $M_{\rm min}$ and is rather mild. Therefore, as with the bubble size distribution, the ionization redshift relative to its average value is nearly independent of the *timing* of reionization. The bottom right panel of Figure 9.3 shows the distribution of reionization redshift in a numerical simulation of the reionization process, illustrating the large dispersion of reionization times.

Because the overdensity distribution narrows as R increases, the typical deviation δz decreases with R; however, the light-crossing time *increases* with R. Thus, there is a critical size above which photons from the far edge of a bubble reach the observer only after the near edge of the bubble has been fully ionized. This critical size then determines the maximum *observable* size. With the presently favored cosmological parameters, this value ≈ 10 comoving Mpc and is nearly independent of the time at which redshift occurred.[10]

9.4.1 The Formal Solution of the Linear Barrier Problem

Before turning to more sophisticated models of reionization, we pause briefly here to derive the mass function of equation (9.21), which will illustrate explicitly how the excursion set formalism allows us to approach these problems. Recall from §3.4.1 that we treat the problem as diffusion in density space, associating a given trajectory with a "halo" (here, an ionized bubble) of mass m_b if the trajectory crosses an absorbing barrier (here specified by δ_B) *for the first time* at mass scale m_b. To solve the diffusion problem, we imagine smoothing the density field around a given point on progressively smaller scales, "zooming in" on that point. If we perform the smoothing such that each stage adds

additional Fourier modes of progressively smaller amplitude, we can replace our mass variable m_b with

$$\sigma_k^2(m_b) = \int_0^K \frac{k^2\,dk}{2\pi^2} P_{\rm lin}(k), \tag{9.27}$$

where the cutoff wavenumber amplitude $K = K(m_b)$. This is a sharp k-space filter; it is convenient because each step in the "zooming in" corresponds to simply adding in more and more k-modes. As each is independent (at least during linear evolution and in the standard cosmological paradigm of Gaussian initial perturbations), each step in our random walk, from m_1 to $m_2 < m_1$, is uncorrelated with previous steps, with standard deviation $\Delta\sigma_k^2 = \sigma_k^2(m_2) - \sigma_k^2(m_1)$.

Unfortunately, a sharp k-space filter has unappealing properties in *real* space, where these halos or bubbles reside (and where we try to observe them): the corresponding real-space window function has oscillatory contributions from large distances and is not confined to a limited spatial region. From this standpoint, sharp filtering in real space is much more appealing, because it describes the well-localized halos or bubbles we desire, and that is what the usual variance of the density field, $\sigma^2(m)$, describes. The fundamental sleight of hand of the excursion set approach is to ignore this distinction by using $\sigma^2(m)$ in place of σ_k^2—a stratagem that is ultimately justified by its utility in matching the results of more detailed numerical simulations. In the remainder of this section, we let $S = \sigma^2(m)$ for notational convenience.

To describe the evolution of fractional overdensity δ with S, we note that the probability of a transition from δ_1 to $\delta_2 = \delta_1 + \Delta\delta$ between S_1 and S_2 is

$$p(\delta_2, S_2)\,d\delta_2 = \Psi(\Delta\delta, \Delta S)\,d(\Delta\delta), \tag{9.28}$$

where

$$\Psi(\Delta\delta, \Delta S)\,d(\Delta\delta) = \frac{1}{\sqrt{2\pi\,\Delta S}} \exp\left(-\frac{(\Delta\delta)^2}{2(\Delta S)}\right) d(\Delta\delta), \tag{9.29}$$

and ΔS is the difference in the variance of the density field between these two scales.

The probability distribution of δ after taking such a step is therefore

$$p(\delta, S + \Delta S) = \int d(\Delta\delta)\,\Psi(\Delta\delta, \Delta S)\,p(\delta - \Delta\delta, S), \tag{9.30}$$

where $p(\delta - \Delta\delta, S)$ is the probability distribution of the density before taking the step, and Ψ is the probability of taking the proper step to reach the final δ. Assuming small step sizes, we can expand both sides in a Taylor series to second order in $\Delta\delta$ (which is first order in ΔS). The left-hand side is straightforward; the right-hand side gives

$$\int d(\Delta\delta)\,\Psi(\Delta\delta, \Delta S)\left(p(\delta, S) - \frac{\partial p}{\partial \delta}\Delta\delta + \frac{1}{2}\frac{\partial^2 p}{\partial \delta^2}(\Delta\delta)^2\right) = p(\delta, S) + \frac{\Delta S}{2}\frac{\partial^2 p}{\partial \delta^2} \tag{9.31}$$

And thus the "evolution" equation for the fractional overdensity becomes

$$\frac{\partial p}{\partial S} = \frac{1}{2}\frac{\partial^2 p}{\partial \delta^2}, \tag{9.32}$$

which takes the form of a diffusion equation.

The boundary conditions are straightforward to describe. First, we know $\delta = 0$ when $S = 0$, by definition, or in other words, $p(\delta|S = 0) = \delta_D(\delta)$, a Dirac delta function. Second, we need to identify trajectories that strike our barrier δ_B with bubbles of the corresponding mass. This is an absorbing barrier in a diffusion problem, as it completely removes any trajectory that reaches this threshold from continuing its random walk. Therefore, $p(\delta_B, S) = 0$.

There are a variety of methods for solving this problem, including Laplace transforms,[11] path integrals,[12] and the method we illustrate here, simple separation of variables. Let us focus on a linear barrier, $B = B_0 + B_1 S$, and transform it to a variable $y = B_1(\delta - B_1 S)$, such that equation (9.32) becomes

$$\frac{\partial p}{\partial S} = \frac{B_1^2}{2}\left(\frac{\partial^2 Q}{\partial y^2} + 2\frac{\partial Q}{\partial y}\right), \tag{9.33}$$

and our absorbing barrier boundary condition becomes $p(B_1 B_0, S) = 0$. We then seek solutions of the form $p(y, S) = g(y) f(S)$. Equation (9.33) does indeed separate, and if we write the separation constant as λ, it is easy to show that the solutions are $f(S) = \exp(\lambda S)$ and $g(y) = \exp[(-1 \pm iX)y]$, where $X = -i(1 + 2\lambda/B_1^2)^{1/2}$. The absorbing barrier boundary condition demands that the solution vanish at $y = B_1 B_0$, which fixes the oscillatory component of g. Thus,

$$p(y, S) = \int_0^\infty dX\, h(X) \sin[X(y - B_0 B_1)] \exp\left[-y - \frac{B_1^2}{2}(1 + X^2)S\right], \tag{9.34}$$

where $h(X)$ represents the amplitude of each mode. This function can be fixed by comparison with the other boundary condition: it is customary to recast this as

$$p(\delta|S = 0) = \lim_{S\to 0} \frac{1}{\sqrt{2\pi S}} \exp\left[-\frac{\delta^2}{2S}\right]. \tag{9.35}$$

If we replace δ with y, it is easy to see that $h(X) = (B_1/\pi)\sin(XB_0B_1)$, which allows us to integrate equation (9.34) and so obtain an explicit expression for the probability distribution,

$$p(\delta, S) = \frac{1}{\sqrt{2\pi S}}\left\{\exp\left(-\frac{\delta^2}{2S}\right) - \exp\left[-\frac{|2(B_0 + i\sqrt{B_0 B_1 S}) - \delta|^2}{2S}\right]\right\}. \tag{9.36}$$

Although the "image" approach described in §3.4.1 does not work here, this equation shows that the final distribution has an analogous form, as the first part is simply the distribution without the absorbing barrier, and the second is the trajectories that have been removed by the absorbing barrier.

difficult to see minus sign before fraction / PE

The probability that a trajectory crosses the barrier in the interval $(S, S+dS)$ is the rate at which these trajectories disappear from the unabsorbed set, or

$$p_{\rm cross}(S) = -\frac{d}{dS}\int_{-\infty}^{B_0 B_1}\frac{dy}{B_1}\,p(\delta, S) = -\frac{B_1}{2}\frac{\partial p}{\partial y}\bigg|_{-\infty}^{B_0 B_1} = \frac{B_0}{\sqrt{2\pi S^3}}\exp\left[-\frac{B(k)^2}{2S}\right], \tag{9.37}$$

where in the second part we have used the diffusion equation (9.33) and the absorbing boundary condition. Finally, the number density of bubbles in equation (9.21) is simply

delete] / PE

$$n_b(m) = \frac{\bar{\rho}}{m}\,p_{\rm cross}(S)\left|\frac{dS}{dm}\right|. \tag{9.38}$$

The technique described here can also be used to solve the halo mass function problem (where B is independent of mass or S), as well as many other interesting problems within the excursion set formalism.

9.5 Recombinations Inside Ionized Regions

lc / PE

Incorporating inhomogeneous recombinations into the excursion set model for ionized bubbles is relatively straightforward. Each H II region obviously contains density fluctuations. Because the recombination rate increases as $(1+\delta_{\rm nl})^2$, where $\delta_{\rm nl}$ is the fully nonlinear fractional overdensity, dense clumps will remain neutral—and optically thick—longer than voids will.

We begin with the simple ansatz that within each ionized bubble there exists a threshold overdensity δ_i below which gas is ionized and above which it is neutral. Any ionizing photons striking these dense blobs—which correspond to LLSs in the postreionization Universe (see §4.4.1)—will be lost to recombinations in the neutral gas and hence are useless for increasing the filling factor of the ionized bubbles. In other words, for an H II region to continue growing, the average separation of these dense blobs must exceed the radius of the bubble. Given a model for the volume-averaged IGM density distribution, $p_v(\delta_{\rm nl})$, we can estimate δ_i by requiring the mean free path between such regions to equal the bubble radius. Clearly, this threshold must increase as the bubbles grow, so that progressively denser gas is ionized with time. Figure 9.6 illustrates this process.

However, ionizing more deeply into the dense gas will also increase the recombination rate per proton, which is

delete subscript v x2 / AA

$$\begin{aligned} A_{\rm rec} &= \alpha(T)\bar{n}_e(1+\delta)\int_{-1}^{\delta_i} d\delta_{\rm nl}\,p_v(\delta_{\rm nl})(1+\delta_{\rm nl})^2 \\ &\equiv \alpha(T)\bar{n}_e C(\delta, R_b), \end{aligned} \tag{9.39}$$

where $C(\delta, R_b)$ is the *local* clumping factor within a bubble of radius R_b and mean overdensity δ, and where we assume that the bubbles are large enough

Note to Karen: I've marked these as AA, but it looks like Barbara misinterpreted my instructions in the copyedit (which were to delete subscript v throughout)

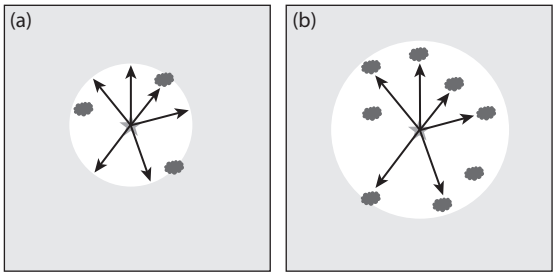


Figure 9.6 Illustration of the role of optically thick regions (or LLSs) in reionization. These always exist due to the clumpiness of the IGM, but they become important only in the late phases of the process. *(a)* Early in reionization, when the ionized regions are small, most lines of sight intersect the edge of the ionized bubble rather than an LLS. Most ionizing photons therefore contribute to increasing the filling factor of ionized gas, and LLSs play only a minor role in slowing the process during its early phases. *(b)* However, once the bubble size grows larger than the mean separation of LLSs, most ionizing photons are absorbed by these objects rather than growing the ionized bubble. The embedded sources therefore no longer contribute to reionization—and so far as they are concerned, the process is over *locally*, even if other regions remain largely neutral.

for linear theory to suffice in this average.[v] The bubble can grow only if ionizing photons are produced more rapidly than recombinations consume them, or in other words, if

$$\zeta \frac{df_{\rm coll}(z, \delta_R, R)}{dt} > \alpha(T)\bar{n}_e C(\delta, R). \tag{9.40}$$

The crucial point is that C depends on both the mean density of the bubble (recall that bubbles correspond to large-scale overdensities) and on its size (through δ_i). Thus, as expected from §9.2.1, inhomogeneous reionization affects the clumping factor. Moreover, the complete model is both "inside-out" on large scales and "outside-in" on small scales. Recombinations become increasingly important as bubbles grow; eventually, they balance ionizations, and the bubble growth saturates in true recombination-limited cosmological Strömgren spheres.

Equation (9.40), which places a constraint on the instantaneous emissivity of ionizing photons, complements our original ionization condition, equation (9.18), which requires that the *cumulative* number of ionizing photons exceed the total number of hydrogen atoms. In reality both conditions must be fulfilled, but in practice one of the two generally dominates. This is essentially because recombinations take over only when δ_i approaches the characteristic density of

delete subscript v x2 / AA

[v] In detail, we actually require the density distribution p_v as a function of large-scale overdensity. Fortunately, in practice, most large ionized bubbles (where recombinations are relevant) are very close to the mean density.

virialized objects, or in other words, when LLSs set the mean free path, as in the lower-redshift Universe (see §4.4.1).

As a consequence, it is possible to combine the two conditions in the excursion set formalism and compute the "bubble" sizes including recombinations. However, this approach requires one conceptual shift: rather than being the actual size of discrete H II regions, the radius R now corresponds roughly to the mean free path of ionizing photons. When recombinations are unimportant, this radius equals the size of bubbles. But once the bubbles "saturate" as Strömgren spheres, neighboring H II regions can touch—only their ionizing photons do not influence one another. This is, in actuality, the same configuration that is present in the postreionization Universe, where ionizing photons are limited by LLSs. The model therefore describes how the "bubble-dominated" topology characteristic of reionization transitions smoothly into the "web-dominated" topology of the postreionization Lyman-α forest, albeit in an inhomogeneous manner across the Universe.

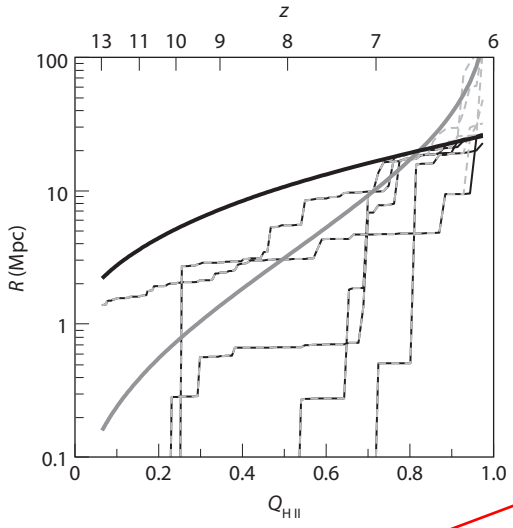
delete subscript v
x2 / AA

The key input parameter is obviously $p_v(\delta_{\rm nl})$, which parameterizes the IGM clumpiness (see also §4.7). In detail, the nonlinear evolution requires cosmological simulations that include coupled dark matter dynamics, gas dynamics, and radiative transfer (to account for the effects of photoheating before and during reionization). This difficult problem has not yet been solved in detail, and so approximate models are generally used. These typically either take the postreionization limit (where the gas is smoothed on the Jeans scale corresponding to a temperature of $\sim 10^4$ K) or appeal to a simple model for structures present before reionization (such as virialized minihalos unable to cool and form stars).

In practice, including recombinations in this manner has a very simple effect: it imposes a *maximum size* on the "ionized regions" that corresponds to the mean free path of an ionizing photon through the inhomogeneous IGM, given the local ionizing background. Bubbles substantially smaller than this limit are almost unaffected by the LLSs, because so few of their ionizing photons strike them.

This picture has important implications for our understanding of the end of reionization. Consider, for example, the evolution of the mean specific intensity of the radiation background, $J \approx \epsilon\lambda/(4\pi)$, where ϵ is the emissivity and λ is the mean free path (see equation 4.44). If we ignored neutral gas inside the ionized bubbles, the mean free path would simply equal the size of the local ionized bubbble, R_b, which, of course, reaches infinity at the end of reionization.

Now, consider how the radiation background grows at a fixed point in the IGM, including inhomogeneous recombinations. When the point is first ionized, J increases rapidly. As the sources inside the bubble ionize their surroundings—gradually adding more sources within the visible "horizon" provided by the bubble edge—J increases slowly, in proportion to R_b. Occasionally, however, the sources ionize a thin wall separating the local bubble from a neighboring H II region. Suddenly, many more sources become visible, and J, along with the local bubble size, instantaneously increases by a large factor. The solid curves in Figure 9.7 illustrate this series of discontinuous jumps in the ionizing background at a few different points in the IGM.



thin / AA

Figure 9.7 Bubble histories for several randomly generated trajectories. The vertical axis shows the bubble radius surrounding a fixed IGM point as a function of the filling factor of bubbles and z; here we arbitrarily fix ζ so that reionization is complete at $z = 6$. Each solid line corresponds to a different IGM point. All include the effect of inhomogeneous recombinations, while the dashed lines ignore them. These differences matter only when the bubbles grow larger than the mean free path of ionizing photons (near the end of reionization), so these are distinguishable only when $Q_{\rm H\,II} > 0.9$. The ~~dotted~~ and ~~dot-dashed~~ lines show the average bubble size R_c and the mean free path λ, respectively, in this model. Furlanetto, S. R., & Oh, S. P., *Mon. Not. R. Astron. Soc.* **363**, 1031 (2005). Copyright 2005 by the Royal Astronomical Society.

heavy dark / light / AA

However, this series of discontinuous jumps cannot continue indefinitely: eventually, the bubble grows large enough that most ionizing photons intercept dense LLSs rather than reaching the edge of the bubbles. From that point, the ionizing background is regulated by the abundance of these systems rather than the global ionized fraction: in effect, the point has reached the "post-overlap" stage even if some of the IGM (at large distances from our point) remains neutral. In Figure 9.7, this situation is illustrated by the range of redshifts (or bubble-filling factors) for which the random trajectories reach λ, where R nearly stops increasing according to this model.

This transformation from an ionizing background regulated by the sizes of H II regions to one regulated by LLSs poses an interesting challenge for studies of reionization. In particular, in the final stages of the process, the mean amplitude of the ionizing background is completely insensitive to the morphology of reionization—and so tells us nothing about that process. Rapid evolution would be indicative not of "overlap" but, rather, of rapid evolution in the absorber population.

9.5.1 The Mean Free Path at High Redshifts

Obviously, the mean free path of ionizing photons plays an extremely important role in regulating the end of reionization. Can we place any constraints on it?

This is a difficult proposition at best. Extrapolating observations at $z < 6$ (equation 4.49) implies that $\lambda \sim 7$ (1) proper Mpc at $z \sim 6$ (10); simple theoretical models predict values in this range as well. However, as the Universe becomes denser and as the ionizing background declines, the densities required to host an optically thick system approach the mean cosmic density. It is therefore not at all clear that such an extrapolation is justified.

For example, equation (4.51) tells us the density of an LLS in terms of the ionizing background. We can make a simple estimate of this background for a stellar population in the context of our simple reionization model. The proper emissivity (in erg cm^{-3} s^{-1}) is

$$\epsilon \sim \zeta \frac{h\nu_{\rm H\ I}\bar{\rho}_b}{m_p}\frac{df_{\rm coll}}{dt}, \tag{9.41}$$

where $\nu_{\rm H\ I}$ is at the ionization edge of H I. This yields an ionization rate

$$\Gamma \sim \epsilon\frac{\lambda\sigma_{\rm H\ I}}{h\nu_{\rm H\ I}} \sim 2.5\times10^{-14}\left(\frac{\lambda}{\rm pMpc}\right)\left(\zeta\left|\frac{df_{\rm coll}}{dz}\right|\right)\ {\rm s}^{-1}, \tag{9.42}$$

where λ is in proper megaparsecs. In equation (4.51), these fiducial values imply that with $\lambda \sim 1$ proper Mpc at $z \sim 10$, $\delta_{\rm LLS} \sim 1$. Thus, LLSs would consist of gas very near the mean density—presumably with physical properties much different from those of the dense LLSs in the moderate-redshift Universe. In fact, more detailed models that attempt to self-consistently match mean free paths of this order with IGM patches find that absorbers must lie inside weakly overdense regions.[13]

A second concern is that the ionizing background—and hence the location of LLSs—fluctuates across the Universe, even discounting the contrast between predominantly ionized and neutral regions. Within bubbles smaller than this mean free path, $\Gamma \propto R_b$ because the volume available for ionizing sources scales as R_b^3, while the flux from each scales as R_b^{-2}. Thus the wide variation in bubble sizes shown in Figure 9.5 will translates into an equally wide variation in Γ. However, as we have argued previously, the increased number of LLSs in small bubbles does not substantially affect the morphology of reionization. Moreover, even within ionized bubbles, Γ has substantial (and systematic) fluctuations as these bubbles expand into low-density regions devoid of sources—although, of course, such regions also have fewer dense blobs capable of becoming LLSs. In practice, once the ionizing background declines to near the cosmic mean, the Γ fluctuations are more important than those in the matter density, so the optically thick systems cluster near the edges of ionized bubbles, where Γ is small.

Figure 9.8 shows some examples of this phenomenon for a series of bubble-filling factors in a numerical simulation. The upper panel shows the inhomogeneous ionizing flux background, which varies by about an order of magnitude within the ionized bubbles, while the lower panel shows the locations

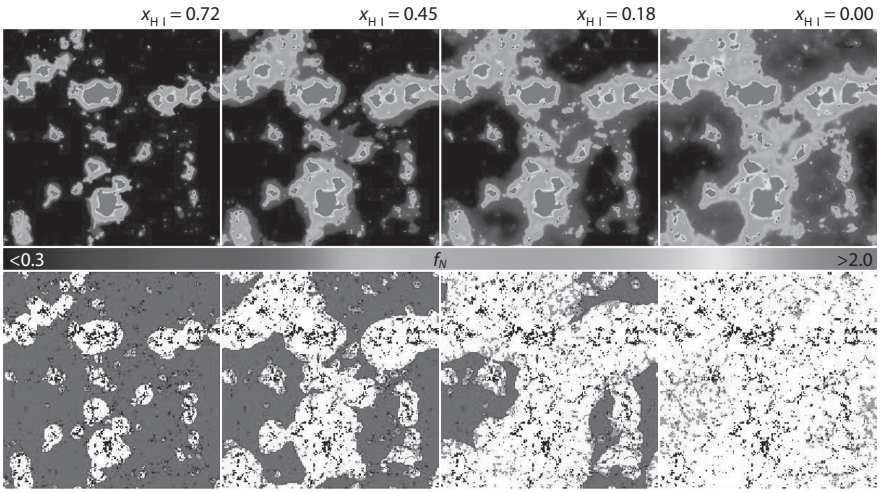


Figure 9.8 IGM absorbers in a "seminumerical" simulation of reionization (see *Color Plate 17* for a color version of this figure). In each panel, we set (by hand) the mean free path of ionizing photons to 10 comoving Mpc; the four columns show different H II fractions at a fixed redshift $z = 10$. The upper panels show the ionizing flux (arbitrary normalization) inside ionized zones according to the color scale. The lower panels show the locations of halos (dark points inside the white regions) and absorbers (lighter points). Crociani, D., et al., *Mon. Not. R. Astron. Soc.* **411**, 289 (2011). Copyright 2011 by the Royal Astronomical Society.

of ionizing sources (dark points) and absorbers (gray points) within the white ionized regions. Note that the absorbers tend to cluster near the edges of the ionized bubbles, even though the underlying density is presumably relatively small there. This is because the ionizing background is weaker in these regions, which are far from the luminous sources.

A final concern is the uncertain amount of small-scale structure in the high-redshift IGM, which depends sensitively on the Jeans mass of this gas and hence the IGM temperature evolution. If, for example, the IGM is not significantly heated before it is ionized, the gas will be much clumpier than in the postreionization Universe, which will render extrapolation from observations useless. We discuss these issues further in §9.9.

9.5.2 Maintaining Reionization

A related question (and one that existing observations can begin to answer) is whether a particular set of ionizing sources can keep the IGM ionized at a sufficiently high level. On a global scale, this requires balancing the recombination rate per unit volume with the emissivity (by number) of ionizing photons, $\dot{n}_{\rm ion}$:[14]

$$\alpha(T) C \bar{n}_{\rm H} \bar{n}_e = \dot{n}_{\rm ion}. \tag{9.43}$$

Unfortunately, this equation has the same ambiguities we have already emphasized. The recombination coefficient depends on the nature of the absorbers (whether case A or case B) as well as the underlying gas temperature (this introduces factor of 2 uncertainties). Moreover, the effective clumping factor C depends on the degree to which dense regions are ionized and is somewhat degenerate with the number of ionizing photons they consume; in detail, it actually depends on the emissivity itself, which sets the self-shielding threshold. An additional difficulty is the implicit assumption that ionizing photons are absorbed instantaneously (or, equivalently, that the elapsed period between emission and absorption is much smaller than both the Hubble time and the characteristic source evolution timescale).

Nevertheless, this equation provides a simple qualitative guide for gauging whether a source population may be able to maintain the observed ionization rate in the Universe. The canonical relation for the required comoving star formation density in galaxies is

$$\dot{\rho}_\star \sim 0.003 f_{\rm esc}^{-1} \left(\frac{C}{3}\right) \left(\frac{1+z}{7}\right)^3 \; M_\odot \; {\rm yr}^{-1} \; {\rm Mpc}^{-3}. \tag{9.44}$$

However, converting the critical rate of ionizing photon production $\dot{n}_{\rm ion}$ to a star formation rate introduces a new set of uncertainties. One substantial difficulty is the escape fraction $f_{\rm esc}$, which is uncertain to at least an order of magnitude. Others are the initial mass function (IMF) of stars, because only the most massive stars produce ionizing photons, and stellar parameters like the metallicity and binarity (see §8.8); the relation here assumes a Salpeter IMF and solar metallicity, both of which are likely conservative and so *overestimate* the required $\dot{\rho}_\star$. Thus, without additional observational constraints on the source populations, equation (9.44) provides only a rough guide.

To ionize most of the IGM in the first place, the cumulative population of stars needs to produce at least one ionizing photon per hydrogen atom in the Universe. Under the same assumptions about the IMF and metallicity as used previously, this condition implies a minimum comoving density of stars after reionization of

$$\rho_\star \sim 1.6 \times 10^6 f_{\rm esc}^{-1} \; M_\odot \; {\rm Mpc}^{-3}. \tag{9.45}$$

Note that this constraint does not involve the clumping factor, since both the number of sources and number of atoms scale the same way with volume.

9.6 Simulations of Reionization

So far we have discussed simplified analytic models of the reionization process. Such models ignore a large number of physical effects, including *(i)* the complexities of radiative transfer, such as shadowing of radiation by a dense absorber; *(ii)* the detailed geometry of the "cosmic web" and source distribution, which is poorly approximated by spherical averaging; *(iii)* the (possible) presence of high-energy photons that can propagate large distances through neutral

gas; *(iv)* the feedback of photoionization and photoheating on the sources of reionization and on the IGM; and *(v)* the nature and clustering of the dense absorbers. It is therefore important to develop more sophisticated numerical approaches to reionization.

In this section we focus on the application of numerical simulations to reionization, with some more specific comments on the numerical implementation of radiative transfer. We refer the reader to §3.7 for more information on the algorithms used in the gravitational and hydrodynamic components of the calculations.

9.6.1 Radiative Transfer Simulations

One option is a full cosmological simulation that attempts to incorporate all the relevant physics, including gravitational dynamics, hydrodynamics, and radiative transfer. This approach is crucial for understanding many of the preceding issues—particularly those involving feedback of reionization itself on the gas distribution. However, including all these effects imposes daunting requirements on the simulations. Most important, we have seen that the relevant scales during reionization easily reach tens of megaparsecs, so simulating a characteristic volume requires a box that spans > 100 Mpc. Additionally, the source halos (even discounting molecular hydrogen cooling) have masses $M_h \sim 10^8\ M_\odot$. Spanning both these scales—with at least 100 particles per galaxy—requires a dynamic range of $\sim 10^{11}$ (in mass), which is very difficult to achieve at present.

As a result, simulations with hydrodynamics—the most difficult of these three physics components to resolve over large dynamic ranges—typically focus on details of reionization that appear on small physical scales, such as feedback on small IGM clumps and the escape of ionizing photons from the local environment of their sources. These sorts of simulations have shown that ionization around galaxies is often highly anisotropic, owing to the dense filaments along which galaxies sit, that photoheating feedback efficiently destroys the smallest gravitationally bound clumps of baryons (or *minihalos*), and that this same feedback moderates the clumping factor throughout the IGM.[15] Simulations like these cannot, however, describe global quantities like the average evolution of the ionized fraction or radiation background, simply because the simulated volumes are too small to include more than one growing ionized bubble.

However, pure gravitational simulations with this dynamic range are relatively straightforward, and radiative transfer optimized for reionization by stellar sources (in which simply following the fate of mono-energetic ionizing photons is not a bad approximation) is relatively simple. Thus, most work to date has focused on dark matter simulations that assume a simple relation between the baryons and dark matter and apply radiative transfer to the resulting baryon field. These simulations very effectively address the detailed geometry of the sources and cosmic web and can at least approximately address the complexities of radiative transfer and the propagation of high-energy photons, but they cannot determine how reionization feedback affects the sources or the IGM (since these are, by definition, hydrodynamic effects).

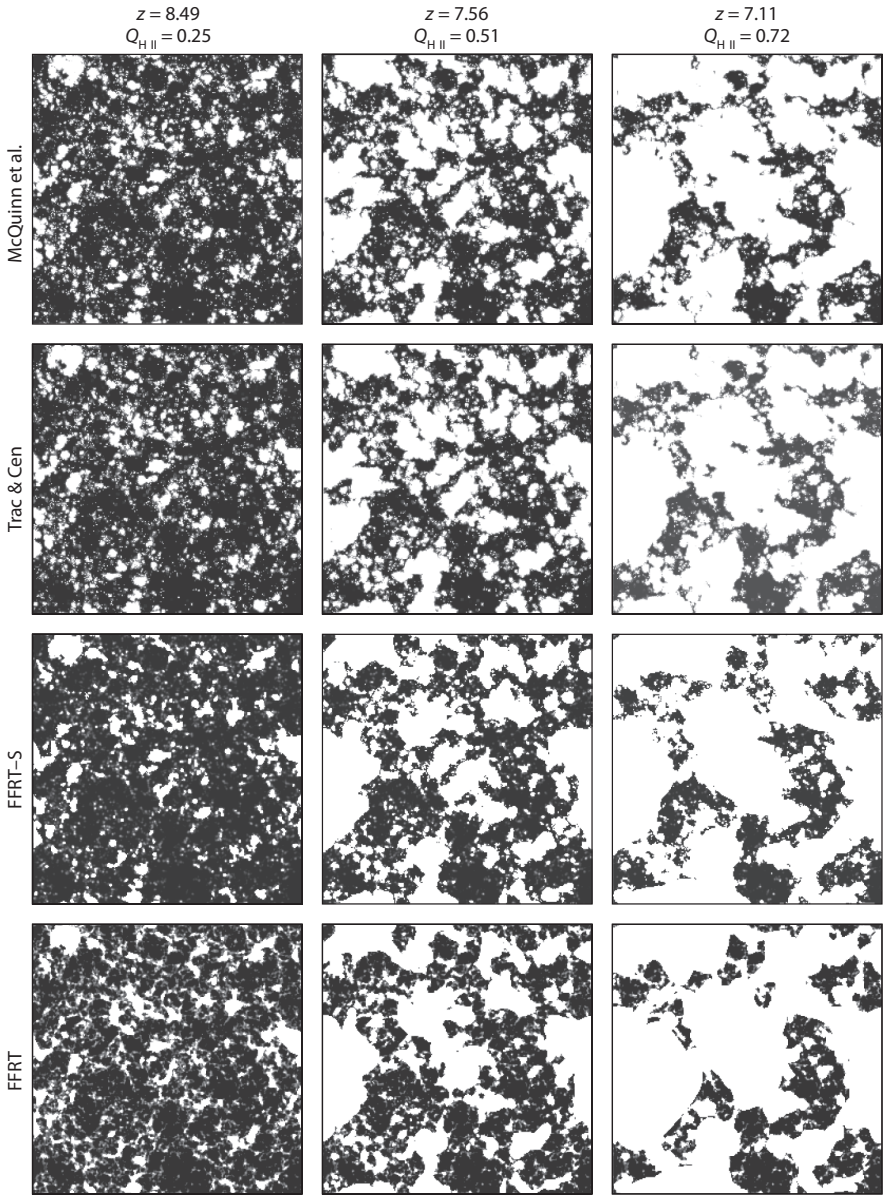


Figure 9.9 Comparison of radiative transfer and seminumerical models of reionization. The three columns show three different times during reionization, with the filling factor of ionized bubbles ~~(here labeled Q_{H})~~ of 0.25, 0.51, and 0.72. The top two rows show two different radiative transfer schemes (both based on adaptive ray tracing). The bottom two rows use seminumerical schemes: the one labeled "FFRT" uses the analytic excursion set model to predict the halo abundance, while the one labeled "FFRT-S" uses the simulated halo field itself. All four rows use exactly the same simulation volume; note the

delete and close up / AA

A variety of radiative transfer algorithms appear in the literature, and fortunately they seem to converge reasonably well in most circumstances.[16] The general problem is very difficult, as computing the specific intensity $I_\nu(t, \mathbf{x}, \mathbf{n}, \nu)$ requires solving a seven-dimensional problem: time t, position $\mathbf{x}$, frequency ν, and direction of propagation $\mathbf{n}$. Furthermore, simulations can contain hundreds of thousands of sources, even excluding the diffuse light generated by IGM recombinations. Thus the complete problem is prohibitively expensive to simulate, and approximate schemes are necessary.

Because each of the many sources illuminates its surroundings over 4π steradians, the number of photon rays that must be included in a calculation is much larger than the number of sources. Codes typically take one of three approaches: *(i)* a Monte Carlo algorithm, in which a large number of photon packets are cast from the sources; *(ii)* adaptive ray tracing, in which a small number of rays are initially cast from each source, spawning new ones as necessary to maintain the desired angular resolution, or *(iii)* a field-based approach, in which photon propagation is abstracted into a continuous field. The first approach is straightforward but faces the most serious convergence challenges. The second most clearly reflects the physics of the problem but is the most challenging technically. Field-based approaches are the fastest but can suffer from unusual artifacts when detailed radiative transfer effects (such as shadowing) become important; they are subject to unphysical diffusion effects in such cases. Nevertheless, the different codes agree rather well. The upper panels of Figure 9.9 compare two different reionization codes (both using the adaptive ray tracing technique) executed within an identical dark matter simulation. The results are clearly very similar.

A second question is how much to specialize the code to the particular problem of reionization. For example, the algorithm can either explicitly incorporate multifrequency sources or focus only on counting ionizing photons. The latter is clearly significantly faster, but the former allows for nonstellar sources and is necessary to trace photoheating accurately. Similarly, in many astrophysical contexts (including LLSs) the ionizing photons emitted during recombinations are important sources, but during reionization such photons are typically absorbed again almost immediately and so get neglected.

Still, even with this sophisticated machinery, numerical simulations are ultimately limited by the same uncertainties that plague analytic models: namely, the physics *inside* high-redshift galaxies is so poorly known that the models are descriptive but not predictive, in the sense that they can accurately predict the statistical properties of reionization given a source model but cannot from first principles generate such a source model.

Figure 9.9 (Continued.) excellent agreement between the radiative transfer schemes and the close match with the seminumerical schemes on moderate and large physical scales. The maps are 143 Mpc/h across and 0.6 Mpc deep. Zahn, O., et al., *Mon. Not. R. Astron. Soc.* **414**, 727 (2011). Copyright 2011 by the Royal Astronomical Society.

A second problem is that these simulations cannot accurately reproduce the properties of photon sinks such as IGM clumping and LLSs, because those depend on the hydrodynamics in and around galaxies as well as on feedback from photoionization. The most sophisticated models prescribe IGM clumping from higher-resolution simulations (together with some assumptions about the distribution of ionized and neutral gas and the relevant level of Jeans smoothing) and/or prescribe the distribution of LLSs based on a semianalytic model.

The most important question is how these numerical approaches compare with the analytic models described earlier. Given all the complexities, the answer—that the analytic models fare extremely well—may be a surprise. Most important, the simulations show large ionized structures, with sizes comparable to those predicted, throughout most of reionization. They confirm that the filling factor of the ionized bubbles, $Q_{\rm H\ II}$, is by far the most important factor in determining the morphology and that the redshift is mostly unimportant. They also show that the clustering of the ionizing sources is the second most important factor and that inhomogeneous recombinations have relatively little effect on the bubble sizes until a threshold H II region size is reached.[17]

9.6.2 Seminumerical Simulations

The general agreement between these disparate approaches has inspired a set of hybrid "seminumerical" algorithms that allow a compromise between the simplicity of the analytic models and the power of a specific realization of reionization.[18] All these approaches follow the same general procedure:

DES: Elsewhere, bulleted lists are set with no extra space between entries. Fix as PE, or allow to avoid reflow?

- First, generate the initial conditions for a cosmological simulation box (usually in a large region of size > 100 Mpc).
- Second, linearly evolve the density field to the desired redshift. As we have seen, doing this is trivial, for the amplitude of density fluctuations simply increases as $D(z)$, independent of scale. Optionally, low-order nonlinear corrections can be applied, such as the Zel'dovich approximation (§4.1).
- Third, identify the source (or dark matter halo) distribution. This is typically done by applying the excursion set approach to the specific density realization of the simulation in one of two ways. One option is to use large cells and compute the expected halo abundance in each one with the analytic excursion set model (using the linear density of each cell as the basis for the conditional mass function of equations 3.43 or 9.19). This is useful for particularly large volumes (>1 Gpc) and/or quick-and-dirty estimates. A second option, useful for more detailed work and/or higher-resolution simulations, is to step through each cell in the simulation volume and smooth the density field on progressively smaller scales, identifying it as a halo whenever it crosses the spherical collapse threshold density (or an improvement on that criterion). This mimics the random walk diffusion process used to generate the halo mass function (see §3.4.1) but applies it point by point to account for real fluctuations in that density field.

The resulting halo field does not match those of numerical simulations exactly but provides an excellent statistical correlation.

DES: Elsewhere, bulleted lists are set with no extra space between entries. Fix as PE, or allow to avoid reflow?

- Finally, generate the morphology of the ionized regions. Again, the density field is smoothed on progressively smaller scales around each pixel, and regions are tagged as ionized if this smoothed field exceeds the excursion set ionization criterion of equation (9.20), that is, if the number of ionizing photons generated within the region (according to some imposed source prescription) exceeds the number of hydrogen atoms.

- Optionally, include a criterion for inhomogeneous recombinations by imposing a maximum bubble size or by weighting the cells according to some estimate of the subgrid clumping and/or self-shielding.

These seminumerical approaches thus represent a fairly direct implementation of the analytic model in specific realizations of the density field. Figure 9.9 shows that the results closely match radiative transfer simulations, at least on large scales. Clearly the broad-brush features are very similar, with ionized bubbles appearing in the same regions and growing to approximately the same sizes in each model. Of course, the detailed shapes of the features are harder to reproduce, especially when two ionized bubbles are near or have just overlapped each other.

Figure 9.10 compares four models in a more quantitative fashion through the power spectrum of the ionized fraction $P_{xx}(k)$, evaluated over the simulated volumes; this is important for many of the observables we will discuss later. At very small scales ($k > 8h\text{Mpc}^{-1}$), the models disagree, largely owing to shot noise in the various prescriptions. On moderate to large scales, the two radiative transfer prescriptions agree extremely well, while the seminumerical prescriptions differ by $\sim$30% late in reionization. These and other statistics show that the hybrid approaches are adequate when accuracy of this order suffices. Most important, the excellent agreement between this implementation of analytic reionization models and the numerical simulations suggests that existing models for the reionization process are quite robust, *given a model for the sources and sinks* (which are themselves almost completely unconstrained by existing observations).

The hybrid approach provides many of the advantages of large-scale simulations (especially the detailed source distribution and cosmic web topology) and with computational costs orders of magnitude smaller. However, it certainly has drawbacks as well. One difficulty is that there is no a priori way to set the excursion set parameters, filtering schemes, and other details of the approach; comparison with simulations has identified the best practical schemes, but the details of the algorithms matter at the $\sim$10% level.[19] Another is that these prescriptions still invoke spherical filtering to paint on the ionization morphology; while the resulting configurations are certainly not themselves spherically symmetric, they do not account for complex radiative transfer effects. Third, the "photon-counting" methods we have studied so far work only for specific

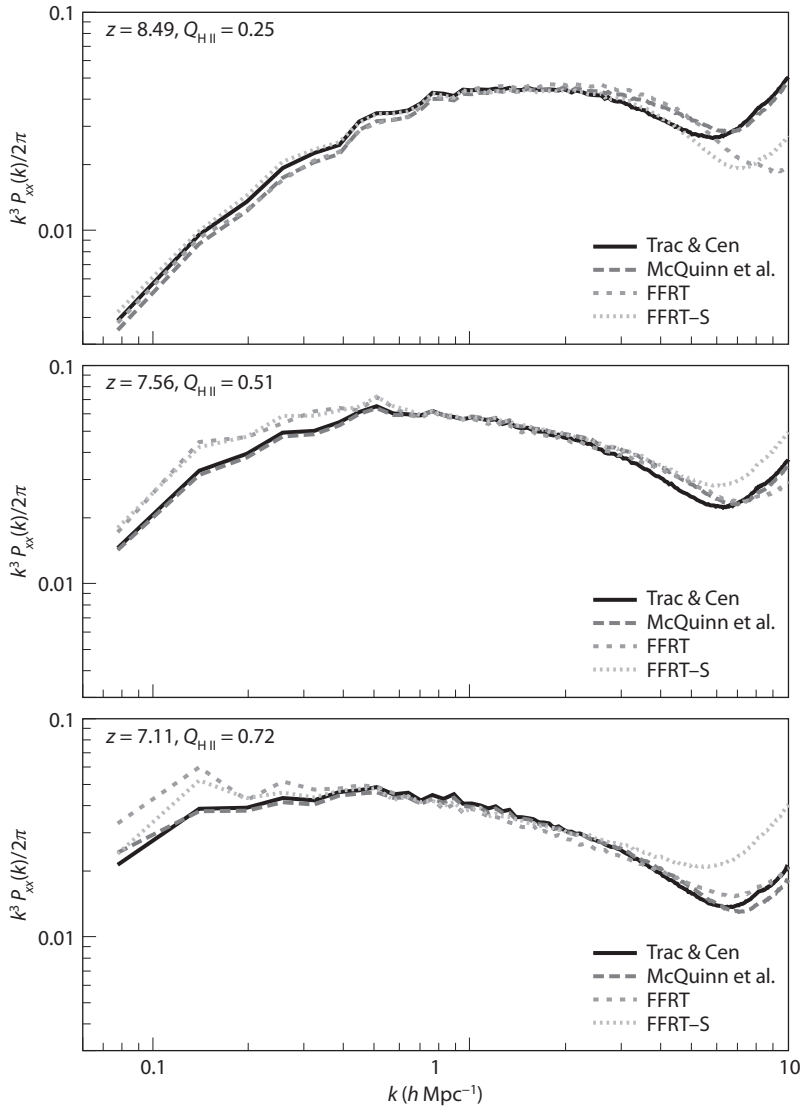


Figure 9.10 Comparison of the power spectrum of the ionization field in the radiative transfer and seminumerical models of reionization. The three panels correspond to the columns of Figure 9.9, and the curves correspond to the four models shown there as well. Note the close match in the predictions of all four models on scales of $k < 8h\mathrm{Mpc}^{-1}$, although the seminumerical schemes do overpredict the power on very large scales in the late stages of reionization. The differences at $k > 8h\,\mathrm{Mpc}^{-1}$ are due to shot noise, which differs between the schemes. Zahn, O. et al., *Mon. Not. R. Astron. Soc.* **414**, 727 (2011). Copyright 2011 by the Royal Astronomical Society.

insert space / PE

classes of sources in which ionizing photons are absorbed shortly after colliding with neutral gas. These schemes have not yet been extended to sources with harder spectra (such as quasars, which we discuss in §9.8).

Perhaps most important, the seminumerical approach cannot be used to follow the progress of reionization through time, because it does not conserve photons. Instead, the global evolution of $Q_{\rm H\ II}(z)$ must be prescribed externally; once that is known, a series of maps can easily be generated, but they cannot then be used to infer anything about the feedback of reionization on the source population, for example. Although $Q_{\rm H\ II}(z)$ in radiative transfer simulations is ultimately determined by imposed source prescriptions as well, they at least allow a self-consistent interaction of the reionization morphology with those sources.

9.7 Statistical Properties of the Ionization Field

Figure 9.10 uses the power spectrum of the ionization fraction to compare the various simulations. The power spectrum offers a convenient way to quantify the statistical properties of a reionization model, and it can be understood intuitively based on the excursion set model of reionization. One relatively rigorous approach to computing the power spectrum on a scale k is to follow two random walks, correlated on all scales $k' < k$, and determine the probability distribution of their fates inside ionized bubbles. This approach provides a reasonably good match to the numerical simulations.[20]

However, we take a simpler, approximate approach here that is informed by the simulation results. We begin by noting that the ionized fraction is not a typical cosmological field, because it is strictly bounded to lie between zero and unity. Thus, we expect the joint probability distribution of the ionized fraction at two different points to take the form

$$\langle x_i(\mathbf{r}_1)x_i(\mathbf{r}_2)\rangle = Q^2_{\rm H\ II} + (1 - Q_{\rm H\ II})\,f(r/R_c), \tag{9.46}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$, and R_c is the characteristic bubble size. Here f is an unknown function containing the physics of the problem, with the limits $f \to 0$ for $r \gg R_c$ and $f \to Q_{\rm H\ II}$ as $r \to 0$. This equation has a simple physical interpretation: if two points are separated by a distance much smaller than the size of a typical H II region, they will either both be ionized by the same bubble, with probability $Q_{\rm H\ II}$, or both be neutral. But if $r \gg R_c$, they must reside in distinct H II regions, so the probability approaches $Q^2_{\rm H\ II}$, with a small enhancement due to the clustering of the bubbles. The correlation function is $\xi_{xx} = \langle x_i(\mathbf{r}_{m1})x_i(\mathbf{r}_{m2})\rangle - Q^2_{\rm H\ II}$ (and the power spectrum is its Fourier transform).

A second restriction on the nature of the correlations arises because of the finite range of the ionized fraction: if $Q_{\rm H\ II} = 1$, every point must be ionized (or $x_i = 1$ everywhere); in that case the correlations must vanish. Thus we need $\xi_{xx} = 0$ when either $Q_{\rm H\ II} = 0$ or 1.

The challenge lies in constructing the function f, which expresses how bubbles encompass two different points separated by a fixed distance. The correlated random walk approach described earlier implicitly computes this factor without treating the ionized patches as discrete objects. We instead use the bubble mass function $n_b(m)$, which necessitates some assumption about their structure. The simplest is, of course, spherical symmetry; unfortunately, this assumption leads to an unphysical suppression in the ionized fraction near R_c. Because the excursion set formalism determines the *maximum* bubble size for which any point is a part, it does not allow for any further overlap of the bubbles. If they are all spherical, it then becomes difficult to pack them in such a way that they ionize all space—this is simplest to see in the limit in which every bubble has the same size, and reionization is then similar to packing a crate with oranges. The gaps between the oranges are impossible to remove in this situation. In reality, of course, the bubbles deform into nonspherical shapes to fill the gap, but that is difficult to model analytically.

We therefore sacrifice rigor to build a simple model that approximates the numerical results.[21] To do so, we split the problem into two regimes. When $Q_{\rm H\,II} < 0.5$, the neutral gaps are large and so are reasonably well modeled by the spherical approximation. Then, taking inspiration from the halo model (§3.6.1), we can explicitly build the joint probability distribution by considering separately (*i*) the probability P_1 that a single bubble ionizes both points and (*ii*) the probability P_2 that the two points are ionized by separate bubbles. In the latter case, we must include the correlations between distinct bubbles. We then have

$$P_1(r) = \int dm\, n_b(m) V_1(m, r) \tag{9.47}$$

$$P_2(r) = \int dm_1\, n_b(m_1) \int d^3\mathbf{r}_1 \int dm_2\, n_b(m_2) \int d^3\mathbf{r}_2 [1 + \xi_{bb}(r|m_1, m_2)], \tag{9.48}$$

where $V_1(m, r)$ is the volume in which the center of a sphere of mass m can lie while simultaneously ionizing two points separated by r, and $\xi_{bb}(r|m_1, m_2) \approx b_{\rm H\,II}(m_1) b_{\rm H\,II}(m_2)\xi(r)$ is the bubble correlation function.

Late in reionization, when $Q_{\rm H\,II} > 0.5$, we set $f = P_1$ in equation (9.46). While this does not include large-scale correlations, by this point the bubbles are so large that the excess correlation on scales beyond the bubble size is negligible. By doing this, we ignore the "two-bubble" term entirely. Thus our expression does not asymptotically approach a form proportional to the dark matter correlation function at late times, $\xi_{xx} \approx \bar{b}^2_{\rm H\,II}\xi$. However, at these late times this limit is reached only at extremely large scales, well beyond the sizes accessible to either observations or simulations. At more moderate scales, the Poisson fluctuations of the discrete bubbles dominate.

Combining these models, we have

$$\langle x_i x_i \rangle (r) = \begin{cases} P_1(r) + P_2(r) & Q_{\rm H\,II} < 0.5, \\ (1 - Q_{\rm H\,II}) P_1(r) + Q^2_{\rm H\,II} & Q_{\rm H\,II} > 0.5. \end{cases} \tag{9.49}$$

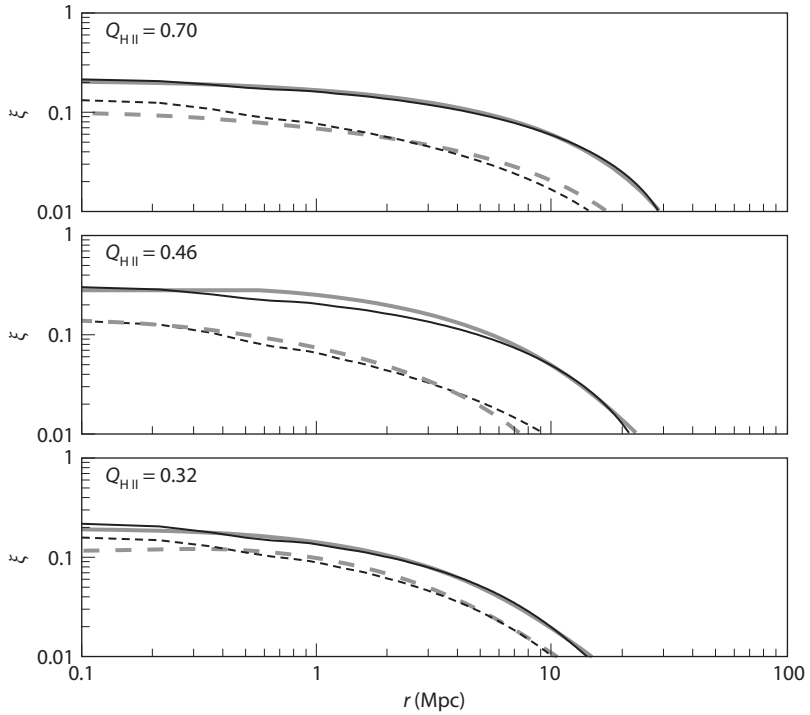


Figure 9.11 Comparison of the autocorrelation function of the ionized fraction (solid curves) and the cross-correlation function of the ionized fraction and density (dashed lines). In each case, the thick lines show our analytic approximations of equations (9.49) and (9.52), while the thin curves show results for a seminumerical simulation in a $100h^{-1}$ Mpc box. The two methods are in quite good agreement at a wide range of ionized fractions. McQuinn, M., et al., *Astrophys. J.* **630**, 643 (2005). Reproduced ~~with~~ permission of the ~~American Astronomical Society~~.

by / AAS. /

The solid curves in Figure 9.11 compare this simple expression with the correlation function found in a seminumerical simulation (including only the linear theory evolution in a $100h^{-1}$ Mpc box) at three different bubble filling fractions. Note the very good agreement at small and moderate scales, which suggests that this simple approach provides good intuition about the properties of the ionization field.

Also of interest is the cross-correlation between the ionized fraction and the underlying density, $\langle x_i(\mathbf{r}_1)\delta(\mathbf{r}_2)\rangle$. Again, it is relatively straightforward to construct a reasonable analytic approximation for this because the excursion set formalism is used for both the halo distribution (which via the halo model describes the density field) and the ionized bubbles. To evaluate it in detail, we can again use some simplifying manipulations. First, suppose that $\mathbf{r}_2$ lies inside the bubble that ionizes $\mathbf{r}_1$ (again, if $\mathbf{r}_1$ is neutral, the product vanishes

automatically). Then, we already know the mean density of the bubble material (equal to the excursion set barrier δ_B in equation 9.20). We can therefore approximate this part of the correlation as

$$P_{\rm in}(r) = \int dm\, n_b(m) V_1(m, r)[1 + \delta_B(m)], \tag{9.50}$$

because the x_i field is unity only inside bubbles, where the mean density is δ_B.

If, however, the point is outside the bubble, we take inspiration from the halo model and assign point $\mathbf{r}_1$ to a halo. We can then approximate the cross-correlation between the bubble and halo as $\xi_{bh}(r) \approx b(m_h) b_{\rm H\ II}(m) \xi(r)$, using the linear theory expression because the distance is large. The contribution from these pairs is

$$P_{\rm out}(r) = Q_{\rm H\ II} - \int dm\, n_b(m) V_1(m, r) + \int dm_b\, n_b(m) \int d^3\mathbf{r}_b\, b_{\rm H\ II} \xi(r), \tag{9.51}$$

where we have used the fact that the mean halo bias is always unity to evaluate the integral over m_h (see equation 3.75). Here the first two terms essentially fix the space available to bubbles to ionize point $\mathbf{r}_2$ without also ionizing $\mathbf{r}_1$; note that the second term in $P_{\rm out}$ cancels the first term in $P_{\rm in}$. The third term contains the correlations. As before, this term is not accurate when $Q_{\rm H\ II}$ is large, because $b_{\rm H\ II}$ encounters difficulty there; however, at these times the bubble radius is so large anyway that the entire correlation term can be ignored. Thus, we have the net approximation

$$\langle x_i(\mathbf{r}_1) \delta(\mathbf{r}_2) \rangle = \begin{cases} P_{\rm in}(r) + P_{\rm out}(r) - Q_{\rm H\ II} & Q_{\rm H\ II} < 0.5, \\ P_{\rm in}(r) - P_1(r) & Q_{\rm H\ II} > 0.5. \end{cases} \tag{9.52}$$

In other words, when $Q_{\rm H\ II}$ is small, we must include correlations from both the bubble at $\mathbf{r}_1$ and from its neighboring bubbles (and, in particular, the excess correlation from their clustering). When $Q_{\rm H\ II}$ is large, we need include only the former effect. Subtracting the P_1 and $Q_{\rm H\ II}$ terms in each case isolates the excess correlations.

Figure 9.11 compares this approximate treatment of the cross-correlation with a seminumerical calculation (thick and thin dashed curves, respectively). Again, the simple model does a rather good job over a range of ionized fractions, though it tends to underestimate the small-scale correlations, because it averages over each bubble.

The important point of this simple model is that the excursion set model reproduces not only the gross properties of the bubble population but also their spatial distribution with respect to the density field. The special nature of the ionization field simplifies many of these calculations, which helps in developing intuitive models that explain the simulation results. Moreover, the correlations can mostly be understood in terms of the average properties of the bubble population, because the individual H II regions are so large that nonlinear effects tend to be washed out.

9.8 Reionization by Quasars and Other Exotic Sources

To this point we have focused on stellar sources of reionization, largely because galaxies seem to dominate the ionizing photon budget at $z \sim 6$. However, quasars present an interesting alternative reionization source, because they have much harder (nonthermal) ionizing spectra than even the hottest stars. Thus, some of their photons can travel much larger distances through the IGM, and the morphology of the ionized and neutral gas is much smoother than the sharply defined H II regions that we have discussed so far.

9.8.1 How Important Are Quasars to Reionization?

There are, unfortunately, very few constraints on the abundance of high-z quasars. The census of very luminous $z \sim 6$ quasars is now fairly well determined, and their abundance seems to decline exponentially at $z > 4$ (see Figure 7.1).[22] Although constraints on the shape of the luminosity function are quite weak, the total ionizing photon emissivity arising from this population of quasars appears to fall a factor of 10–50 short of that required to maintain reionization at that time when we use $C = 3$ and the arguments in §9.5.2.

Nevertheless, it is relatively easy to imagine that much smaller black holes—in particular those characteristic of the small galaxies common at high redshifts—could play an important role in at least partially ionizing the IGM. As discussed in §7.5, at lower redshifts it is now clear that black holes are both ubiquitous and closely related to their host galaxies. For the purposes of a simple estimate we simply scale $M_{\rm BH}$ to the total halo mass M_h, so that the mass density in black holes is $\rho_{\rm BH} = f_{\rm BH} f_{\rm coll} \bar{\rho}_b$. As a fiducial value, we scale $f_{\rm BH}$ to its local value in massive galaxies, $\sim 10^{-4}$, but as discussed in §7.5.1, this value may evolve at high redshifts (and appears to be an underestimate for the brightest $z \sim 6$ quasars).

Now, let us consider how large $f_{\rm BH}$ must be to significantly ionize the IGM. These ionizations come from two sources: primary photoionizations from the quasar photons themselves, and secondary ionizations from the energetic secondary electrons produced after the initial photoionization. For a hard nonthermal spectrum $L_\nu \propto \nu^{-1}$, the latter dominate and deposit (very crudely) a fraction of the energy $f_i \sim x_{\rm H\,I}/3$ in ionizations. If the black holes have a radiative efficiency (relative to their rest mass) ϵ and emit a fraction $f_{\rm UV}$ of their energy above the ionization threshold with a fraction $f_{\rm esc,q}$ of that energy escaping the host galaxy, the expected number of ionizations per hydrogen atom is

$$N_X \sim 0.5 f_{\rm esc,q} \left(\frac{\epsilon}{0.1}\right) \left(\frac{f_{\rm UV}}{0.2}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{f_i}{1/3}\right). \qquad (9.53)$$

Thus, the local black hole–halo relation makes a plausible argument for a substantial contribution of quasars to reionization. Note, however, that the secondary ionizations become less and less common as $x_{\rm H\,I}$ decreases, so lower-energy photons (either from quasars or stars) are still necessary to complete reionization (as discussed in the next section).

Unlike for stars, the escape fraction $f_{\rm esc,q}$ is likely to be quite high for quasars. Because all the quasar ionizing radiation emerges from a single source, it is much more likely to carve transparent channels in the ISM of the galaxy. Moreover, much of the ionizing energy comes from relatively high energy photons that have an easier time traversing their host galaxy without interacting.

The unresolved X-ray background offers a constraint on this scenario, because such a high-redshift quasar population would have produced hard X-rays ($\geq$10 keV) that are still free streaming, with some redshifting into the well-observed soft X-ray band. Approximately 93 $\pm$ 3% of the soft X-ray background has been resolved; the best estimate for the unresolved component is $J_X \sim 0.3$–$1 \times 10^{-12}\,{\rm erg\,s^{-1}\,cm^{-2}\,deg^{-2}}$ in the 0.5–2 keV band.[23]

Suppose that black holes produced the high-redshift X-ray background at a median redshift z and emitted a fraction $f_{\rm HXR}$ of their energy in the [0.5–2](1 + z) keV range. The flux received at Earth would then be $J = (c/4\pi)\rho_{\rm HXR}/(1+z)$, where $\rho_{\rm HXR}$ is the comoving energy density in hard X-rays produced by this early generation of black holes. Thus

$$J_{\rm X} \approx 10^{-13}\, f_{\rm esc,q}^{-1} \left(\frac{f_{\rm HXR}/f_{\rm UV}}{0.2}\right)\left(\frac{1/3}{f_i}\right)\left(\frac{N_X}{0.5}\right)\left(\frac{10}{1+z}\right)\ {\rm erg\,s^{-1}\,cm^{-2}\,deg^{-2}}, \tag{9.54}$$

where the fiducial choices for $f_{\rm HXR}/f_{\rm UV}$ and $\langle E\rangle$ are appropriate for a spectrum with $L_\nu \propto \nu^{-1}$ ranging from 13.6 eV to 10 keV.

Interestingly, this value is comparable to the presently observed unresolved component. Thus, the X-ray background required if quasars *alone* reionized the Universe probably violates observed limits,[24] but they could still have made a substantial contribution to the ionization budget; thus it is certainly useful to consider scenarios in which quasars drove or affected the reionization process. Stellar-mass X-ray binaries could also have contributed to the X-ray background,[25] as we discuss in §12.3.2.

Moreover, it is relatively easy to imagine scenarios in which black hole accretion played a much larger role. One possible way to evade these constraints is with a population of "miniquasars" built from smaller black holes that may have formed through different channels than the very bright observable quasars. In such miniquasars, most of the UV photons may come from an accretion disk, while hard X-rays instead come from synchrotron/inverse-Compton emission. The relative contribution of the two components is extremely uncertain, and if the nonthermal tail is relatively insignificant, the empirical X-ray background constraint will tell us little about the total contribution of black holes to reionization.

9.8.2 Ionized Bubbles around Quasars

The primary difference between quasars, which typically have nonthermal spectra in the UV and X-ray regimes, and stars (which are nearly thermal and so have very few high-energy photons) is that one cannot simply assume that all the ionizing photons are absorbed in a narrow region around the ionization

front; instead, the higher-energy photons can propagate large distances through the intergalactic medium. The comoving mean free path of an X-ray photon with energy E is

$$\lambda_X \approx 11\,\bar{x}_{\rm H\,I}^{1/3}\left(\frac{1+z}{10}\right)^{-2}\left(\frac{E}{300\ {\rm eV}}\right)^{3}\ {\rm Mpc}; \tag{9.55}$$

thus, photons with $E > 1.5[(1+z)/15]^{1/2}\bar{x}_{\rm H\,I}^{1/3}$ keV propagate an entire Hubble length before interacting with the IGM. Many of the soft X-rays therefore escape the ionized bubble but deposit their energy (as ionization and heat) in the surrounding gas, "preionizing" and "preheating" it before the ionization front itself reaches the gas.

In this case where photons leak past the "ionization front" marking the boundary between the mostly ionized and mostly neutral gas, the photon-counting arguments implicit to §9.1 are not sufficient: the assumption of a two-phase medium (highly ionized and completely neutral) breaks down. Instead we must more carefully examine the radiative transfer of ionizing photons through these regions. For simplicity, we consider a model universe composed entirely of hydrogen; including helium complicates the equations but adds no essential new physics. As a photon travels away from its source, it encounters absorption that depends on the local ionized fraction as well as the photon energy. The total optical depth experienced by a photon with frequency ν that has traveled from a source to a radius r is

$$\tau(\nu, r, t) = \int_0^r \sigma_{\rm H\,I}(\nu) n_{\rm H\,I}(r, t)\, dr', \tag{9.56}$$

where $n_{\rm H\,I}$ is the local H I density (which may evolve either through the cosmological expansion or the neutral fraction) and where we have explicitly noted the time dependence, since the ionized region will grow as more and more photons are pumped into it. We have also assumed that $r \ll c/H(z)$, so that we can ignore the cosmological redshift of the photon. The ionization rate at this position is then

$$\Gamma(r, t) = \int_{\nu_{\rm HI}}^{\infty} \frac{d\nu}{h\nu} \frac{L_\nu e^{-\tau(\nu,r,t)}}{4\pi r^2} \sigma_{\rm H\,I}(\nu)\left[1 + \left(\frac{E - E_{\rm H\,I}}{E_{\rm H\,I}}\right) f_i(E - E_{\rm H\,I})\right], \tag{9.57}$$

where $L_\nu = (dL/d\nu)$ is the monochromatic luminosity (per unit frequency) of the source, $E_{\rm H\,I} = 13.6$ eV is the ionization potential of H I, $E - E_{\rm H\,I}$ is the energy of the photoelectron, and $f_i(E - E_{\rm H\,I})$ is the fraction of this energy that goes into secondary ionizations as this electron scatters through the ambient medium. This last factor describes the fate of the high-energy electrons; it is small for photons near the ionization threshold and (very roughly) approaches $f_i \sim x_{\rm H\,I}/3$ at high energies. A comparable fraction of the energy goes into collisional excitation of line transitions; the remainder goes into heating (see further). These fractions have been computed much more precisely using basic atomic physics.[26]

The ionization rate at each position is then governed by

$$\frac{dn_{\rm H\ I}}{dt} = \Gamma n_{\rm H\ I} - \alpha_B(T) n_e n_{\rm H\ II}, \tag{9.58}$$

where (again ignoring helium) $n_e = n_{\rm H\ II} = n_H - n_{\rm H\ I}$. We assume case-B recombination (i.e., local absorption of the recombination photons) for simplicity; otherwise, the radiative transfer equation must include a source function for recombination photons as well. This "on-the-spot" approximation is usually a good one because the recombination photons are emitted near the ionization threshold and so have short mean free paths. Note that we have left the clumping factor C off equation (9.58), because integrating the ionization front evolution over space allows us to include the detailed density profile. However, it can easily be incorporated into the last term to account for clumping below the resolution of the calculation grid.

Roman / PE

Because the recombination rate depends on temperature T (and often because the temperature is of intrinsic interest), we must also trace its evolution with (see also §4.3.1)

$$\frac{dT}{dt} = -2HT + \frac{2T}{3}\frac{d\ln(1+\delta)}{dt} - T\frac{d\ln(2-x_{\rm H\ I})}{dt} + \frac{2}{3k_B n_{\rm tot}}(\mathcal{H}-\Lambda), \tag{9.59}$$

where $\mathcal{H}$ is the total radiative heating rate, Λ is the total radiative cooling rate, and $n_{\rm tot}$ is the total particle number density. From left to right, these terms describe adiabatic cooling due to the Hubble expansion, adiabatic heating or cooling due to local density inhomogeneities, the change in the total particle density due to ionizations and recombinations, and radiative processes.

At high redshifts, radiative heating and cooling are typically dominated by photoheating and inverse-Compton cooling, respectively. The former is

$$\mathcal{H}_{\rm ph} = \int_{\nu_{\rm H\ I}}^{\infty} d\nu\, \frac{L_\nu e^{-\tau(\nu,r,t)}}{4\pi r^2}\sigma_{\rm H\ I}(\nu)\,(E - E_{\rm H\ I})\, f_h(E - E_{\rm H\ I}), \tag{9.60}$$

where $f_h(E - E_{\rm H\ I})$ is the fraction of the photoelectron energy that goes into heating. It is large for photons near the ionization threshold and (very roughly) approaches $f_h \sim 1 - 2x_{\rm H\ I}/3$ at high energies.[27] The Compton cooling rate $\Lambda_{\rm comp}$ is given by (see equation 2.38)

$$\frac{2}{3}\frac{\Lambda_{\rm comp}}{k_B n_{\rm tot}} = \frac{1-x_{\rm H\ I}}{2-x_{\rm H\ I}}\,\frac{(T_{\rm CMB}-T)}{t_c}, \tag{9.61}$$

where $t_c \equiv (3m_e c)/(8\sigma_T u_{\rm CMB})$ is the Compton cooling time (see also §2.2), σ_T is the Thomson cross section, and $u_{\rm CMB} \propto T_{\rm CMB}^4$ is the CMB energy density. The first factor on the right-hand side accounts for energy sharing by all free particles.

Figure 9.12 show some example ionization and temperature profiles around a relatively bright quasar at $z = 10$ with $L_B = 10^9\ L_\odot$. The source is assumed to emit steadily after it turns on, and the different curves take $t = 10^6$, $10^{6.5}$, 10^7, $10^{7.5}$, and 10^8 yr after ignition. The calculation assumes an initial IGM temperature of $T = 10$ K and a uniform IGM at the mean density. As expected, the ionization front sweeps outward over time. Behind it, the gas lies

The other important process is line cooling, which dominates in some particular temperature ranges (see Section 11.6.3). / AA (use Section symbol in place of word)

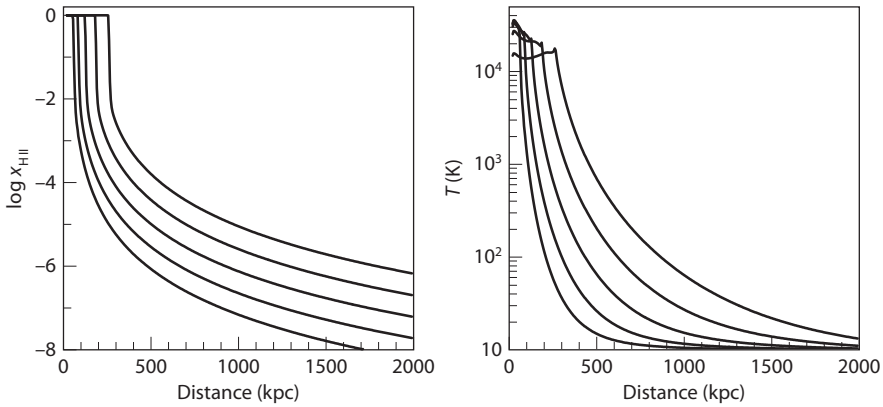


Figure 9.12 Example ionization and temperature profiles around a relatively bright quasar at $z = 10$, with $L_B = 10^9 L_\odot$. The source is assumed to emit steadily after it turns on, and the different curves take $t = 10^6$, $10^{6.5}$, 10^7, $10^{7.5}$, and 10^8 yr after ignition, from bottom to top. The calculation assumes an initial IGM temperature of $T = 10$ K and a uniform IGM at the mean density (including helium); note that distances are measured in proper (not comoving) units.

in ionization equilibrium, with $x_{\rm H\,I} \propto r^2$. The ionization front itself—which we define to be the distance $0.1 < x_{\rm H\,I} < 0.9$—is narrow, but residual ionization (and heating, which can be substantial) extends several megaparsecs from the front itself. The gas leading the front is *not* in ionization equilibrium, as the ionization front continues to sweep outward if the source remains luminous, and the gas outside steadily increases in both temperature and ionized fraction until it reaches the highly ionized limit.

In particular, because the recombination time in this outer region is so long (at least while the ionized fraction itself is small), the relatively low level of heating and ionization contributed by each quasar is cumulative. After many generations of AGN, the gas that remains outside H II regions gradually becomes increasingly ionized, potentially until the ionized fraction saturates at ∼0.5, when secondary ionizations become inefficient.

This gradual ionization and heating of the otherwise untouched gas provides one of the key differences between stellar and quasar reionization. Others are primarily driven by differences in the source luminosities and abundances: to the extent that quasars are rarer and more luminous than star-forming galaxies, they will produce larger, rarer H II regions in the IGM, in which the ionized fraction and density field are less correlated. We discuss some of the observational signatures of these differences in later chapters.

9.8.3 Helium Reionization

So far we have focused purely on the reionization of intergalactic hydrogen. The first ionization potential of helium, 24.4 eV, is sufficiently close to that of hydrogen that helium is almost definitely singly ionized at the same time as

hydrogen. However, stripping the second electron requires 54.4 eV, which is well beyond the blackbody peak of typical hot stars (although very massive metal-free stars can at least partially ionize helium; see §5.4). We therefore expect a significantly different ionization history for He II.

Nevertheless, many of the same tools we have already developed can be used to follow the creation of He III. Helium can easily be incorporated into the formalism of §9.8.2 by adding a multispecies network that traces the evolution of He II and He III. In practice, most high-energy photons are absorbed by He II, but (because helium is relatively rare) the secondary electrons still deposit most of their energy as heat or in ionizing and exciting H I. The ionization and heating profiles (as in Figure 9.12, which does include helium) do not qualitatively change.

Similar calculations for stellar sources show that only very massive metal-free stars can produce He III, although even in optimistic models the He III fraction rarely rises to unity.[28] Moreover, once these stars fade away, the He III rapidly recombines into He II because its recombination time is much shorter than that for H II (see §4.5) and therefore shorter than the age of the Universe t_H,

$$\frac{t^B_{\rm rec,\,He}}{t_H} \approx 0.2 \left(\frac{8}{1+z}\right)^{3/2}. \tag{9.62}$$

Thus, there may have been a brief phase of ionized helium during the cosmic dawn, but it likely ended with the death of these primordial stars.

However, radiation from quasars could provide a more sustained source of high-energy photons. We have already seen that these sources can plausibly ionize hydrogen; can they do the same for He II? The primary difference from our earlier calculation is that fast secondary electrons produced in the ionization process do not efficiently ionize He II, because its collisional ionization cross section is more than an order of magnitude smaller than that of H I (and when hydrogen is fully ionized, the energy loss rate to other electrons is also much more rapid). Without secondary ionizations, the crucial parameter is the mean photon energy per ionization $\langle E_i \rangle$. If $L_\nu \propto \nu^{-1}$ from 54.4 eV to 2 keV (beyond which the IGM is optically thin), this energy is ∼200 eV. Assuming that all the high-energy photons ionize He II rather than H I (e.g., if stellar sources ionize the latter first), we find that the number of ionizations per helium atom could be

$$N_{\rm ion,He} \sim 0.6 f_{\rm esc,q} \left(\frac{\epsilon}{0.1}\right) \left(\frac{f_{\rm UV,He}}{0.1}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{200\ {\rm eV}}{\langle E_i \rangle}\right). \tag{9.63}$$

Here $f_{\rm UV,He}$ is the fraction of the quasar's luminosity emitted above 54.4 eV. Of course, given the rapid recombination time, these early quasars are unlikely to have maintained more than a low level of He III in the IGM.

Despite this estimate, just as for H I, the *observed* high-z quasar population produces far fewer He II–ionizing photons. In fact, estimates based on the measured quasar luminosity function (appealing to both photon-counting arguments and the required emissivity to maintain reionization) predict that

He II reionization must not have begun until $z \sim 3$, near the peak of the quasar era.[29] Indeed, a number of lines of evidence indicate that the event occurred at roughly that time, though none are as yet definitive. We list these efforts here because they make an interesting comparison to the constraints on H I reionization that we discuss later:

- The mean optical depth of the He II Lyman-α forest appears to increase rapidly beyond $z \sim 2.8$.[30] In §4.7 we argued that an apparently similar increase in the H I forest optical depth at $z \sim 6$ could not robustly be interpreted as evidence for reionization. ~~But~~ the case for helium is more ~~secure~~: because the atomic number density of helium is smaller and its recombination rate is faster, its Gunn-Peterson optical depth is only $\tau \sim 14\Gamma^{-1}_{\rm He\ II,-14}(1+z/4)^{9/2}$, where $\Gamma_{\rm He\ II,-14}$ is the He II ionization rate in units of 10^{-14} s^{-1} (roughly the measured value). Thus, He II becomes transparent in the late stages of reionization; moreover, it does *not* have an opaque damping wing that can conceal highly ionized regions (see §11.4). Additionally, He II reionization is accomplished by rare, bright sources whose illumination can create large (many megaparsec) ionized bubbles even before the process is completed. Together, these factors imply that the He II Lyman-α forest is a much cleaner probe of reionization than for H I.

- Moreover, the He II forest shows substantial fluctuations at $z > 2.8$, from being nearly opaque to very transparent.[31] Such regions are difficult to arrange if the IGM is highly ionized, because they require a dearth of quasars over several hundred megaparsecs, which is very unlikely. Unfortunately, the enormous optical depth of the H I forest at $z \sim 6$ masks the analogous fluctuations, and so this test is much more difficult to repeat with hydrogen.

- A number of measurements of the H I forest show a peak in the IGM temperature at $z \sim 3$.[32] The most natural interpretation is photoheating from helium reionization (see §9.9).

- There is some evidence for a hardening in the metagalactic ionizing background at $z \sim 3$, as measured by the ratios of some metal lines. For example, C IV has an ionization potential just above that of He II, while that of Si IV is far above that point. Once He II is ionized and the IGM becomes transparent to photons above 54.4 eV, we expect the abundance of C IV to decrease relative to Si IV as well (see §4.6). Some (but not all) measurements show such an decrease.[33] At $z \sim 6$, the analogous process at the H I edge should show an increase in higher ionization states (e.g., C IV) relative to low ionization states (e.g., O I) (see §4.6). Tentative evidence for such evolution does exist, but the scarcity of metal-line systems at $z > 5$ and their likely positions inside highly overdense systems complicate their interpretation in this case.

persuasive / AA (factual correction)

delete / capitalize / AA

DES: Elsewhere, bulleted lists are set with no extra space between entries. Fix as PE, or allow to avoid reflow?

Clearly, He II reionization is at best an imperfect analog to hydrogen reionization, but it does allow us to test a number of the same ideas—especially those relating to the ionizing background and its interaction with the IGM. In particular, it has the key advantage of occurring at $z \sim 3$, where measurements of the H I Lyman-α forest offer a much clearer picture of the IGM. Helium reionization may therefore offer a useful test bed for understanding the tail end of hydrogen reionization, when LLSs dominate the absorption of ionizing photons, and IGM structure is crucial.

9.8.4 Exotic Reionization Scenarios

It is also possible that much more exotic processes—such as dark matter decay or annihilation, or primordial black hole evaporation—helped (or even completed) the reionization of the IGM. Any such exotic process that produces photons with $E > 13.6$ eV to which the IGM is opaque could have contributed to ionizing (and possibly heating) the IGM. For example, dark matter decay—even with a timescale many times the present age of the Universe—could in principle have reionized the entire IGM, so long as $> 10^{-8}$ of the total rest energy of the dark matter particles went into ionization.[34]

Although such models are quite speculative, they would produce very different patterns of reionization and so are interesting from a phenomenological perspective. For example, dark matter is fairly uniformly distributed at high redshifts, so decay would cause a nearly uniform ionizing background and hence a nearly uniform ionized fraction (moderated only by inhomogeneous recombinations and possible escape of the decay products from the source region). Annihilation would provide a clumpier source distribution but would still cause much smoother reionization than stars or quasars.

9.9 Feedback from Reionization: Photoheating

As described in §4.3.1 and §9.8.2, after photoionization (some of) the excess energy deposited in the photoelectron is transformed to heat through scattering. This heating can be substantial: for a spectrum typical of a star-forming galaxy, $\Delta T \sim 12{,}500$–$30{,}000$ K (see §4.3.1), while for quasar sources, might be $\Delta T \sim 10^5$ K. Owing to X-ray heating, the IGM temperature was rather uncertain before reionization (as we discuss in §12.3.2), but this photoheating almost certainly increased it by nearly an order of magnitude, which has a number of important consequences.

it / AA

9.9.1 Photoheating and the IGM

If reionization were uniform, this dramatic heating would have left the IGM essentially isothermal. However, we have seen that in fact the process was driven by large-scale density fluctuations, with overdense regions (full of galaxies) reionized first and underdense regions (devoid of galaxies) reionized last

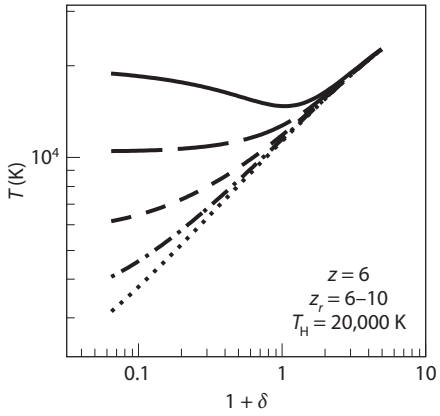


Figure 9.13 IGM temperature–density relation following H I reionization. All curves assume $z = 6$ and take a postreionization temperature of $T_H = 20{,}000$ K. The solid, long-dashed, short-dashed, dot-dashed, and dotted curves set the reionization redshift at $z_r = 6$, 7, 8, 9, and 10, respectively. Furlanetto, S. R., & Oh, S. P., *Astrophys. J.* **701**, 94 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

(see the bottom right panel of Figure 9.3), which translates into systematic IGM temperature fluctuations. Once reionization ended, the rapid photoheating ceased (constrained by the recombination rate within the IGM gas). Thus the overdense regions began cooling *earlier* and had systematically cooler temperatures at the tail end of reionization (see the bottom center panel of Figure 9.3).

Figure 9.13 shows this process quantitatively via the IGM temperature-density relation. We show this relation at $z = 6$ computed from the excursion set reionization model of §9.4 for a variety of scenarios in which reionization ends between $z_r = 6$ and 10 (thus the different curves do not represent a time sequence from one model but, rather, a sequence of different reionization models, with the time of observation held constant). Immediately following reionization (solid curve), the low-density voids are systematically hotter than gas near the mean density, simply because the former were the last to be ionized and so still lie near the postreionization temperature. (Note that overdense gas is hot as well, due to the adiabatic heating from ongoing structure formation.)

This kind of *inverted* temperature–density relation is strongly characteristic of "inside-out" reionization, in which large-scale overdensities are ionized first and underdense voids last. Inside-out models are generic to stellar reionization, because its morphology closely traces the underlying cosmic web. (Note, however, that this does not mean that *small-scale* overdensities are ionized last—in fact these LLSs typically maintain relatively large neutral fractions until the very late stages.) If, however, rare, luminous sources (such as quasars) drove reionization, the ionized bubbles correlate less strongly with the density field, and the associated temperature inversion weakens (or even disappears—as is likely

to be the case with helium reionization at $z \sim 3$). Thus, the IGM temperature–density relation provides a good test of the morphology of reionization.

The expansion of the Universe causes gas at all densities to cool adiabatically. However, because underdense voids expand more rapidly than average, this cooling occurs fastest at low densities, gradually erasing the initial (inverted) temperature–density relation. Because adiabatic cooling occurs over an expansion time, the characteristic cooling timescale is the Hubble time (see the first two terms on the right-hand side of equation 9.59). Thus, the interesting observational signature of inside-out reionization fades after a relatively short time, and the temperature–density relation approaches the universal asymptote in which photoheating following recombinations balances the adiabatic cooling.

Of course, reionization was also stochastic, with regions of a given density having many different reionization histories (driven by the nearby halo population). Thus, the temperature-density relation is imperfect, with scatter of $\sim$30% at a given density. This scatter (and its dependence on density) also depends on the reionization model; rarer sources induce more scatter.

Photoheating from reionization not only increased the IGM temperature but also affected its structure: the accompanying thermal pressure increased the effective Jeans mass ($M_J \propto T^{3/2}$; see equation 3.15), evaporating existing small-scale structures and preventing accretion onto small dark matter clumps. In the diffuse IGM, this effect is usually interpreted as a decrease in the clumping factor C. When M_J was small, before reionization, very small dark matter halos could retain their baryons, producing a great deal of small-scale structure, while after reionization, these structures evaporated and C decreased. Fortunately, this smoothing was relatively insensitive to the precise postreionization temperature, because (in most models) any reasonable amount of photoheating had already increased the temperature by a very large factor.

However, following this evolution in detail requires quite sophisticated numerical simulations that (a) resolve the small-scale IGM structure and (b) include coupled radiative transfer and hydrodynamics. To date this has been possible only in relatively small-volume simulations that do not fully account for the large-scale morphology of reionization; fortunately, the insensitivity of the resulting clumping evolution to the details of the reionization process suggests that these results—in which the clumping factor can decrease by nearly a factor of 2 owing to photoheating—are robust.

Directly observing photoheating is a challenge, especially at very high redshifts, in which the Lyman-α forest is nearly saturated in absorption. In addition to the Jeans smoothing itself (which affects small-scale power in the forest), heating also increases thermal broadening in the line profiles. These manifest themselves in both statistical measures of the forest (like the power spectrum, where small-scale structure is erased) and in the lines (where spectral broadening leaves less curvature in the spectrum). Although these techniques have not yet been feasible at high redshifts, they are easier at moderate redshifts ($z \sim 3$) around the time of He II reionization and have been extensively applied there. Both methods have provided measurements of the evolution of the mean

temperature with redshift and show heating at $z \sim 3$, of roughly the magnitude expected if quasars were responsible for the event (see §9.8.3).

However, these methods have not yet offered strong constraints on the temperature–density relation, primarily because the forest is mostly sensitive to only a narrow range of densities at any one redshift. One interesting way to avoid this problem and extend it to high redshift is by comparing constraints from multiple Lyman lines. With their weaker oscillator strengths, Lyman-β and Lyman-γ sample different parts of the density field (see Figure 4.16); comparing their different optical depths as a function of redshift may therefore reveal how the temperature evolved following reionization.

9.9.2 Photoheating and Virialized Objects

Photoheating affects not only diffuse IGM gas but also gas inside virialized objects. If such a halo has $T_{\rm vir} < 10^4$ K, photoionization will heat the gas above the escape velocity of the halo, allowing the baryons to evaporate. Moreover, once IGM gas is heated, it will ignore small dark matter potential wells, preventing the accretion of gas onto existing galaxies and suppressing subsequent star formation.

The Jeans mass (or, more properly, the filter mass) in the IGM is $M_J \sim 10^5\ M_\odot$ if the gas simply cools adiabatically after decoupling from the CMB (see §3.2). This is far below the atomic cooling threshold ($T_{\rm vir} \sim 10^4$ K corresponds to $\sim 10^7\ M_\odot$; see §3.3), so although these dark matter clumps can accrete baryons, they cannot go on to form stars. Instead, they remain as dense clumps sprinkled through the IGM. Moreover, because the mass function is so steep at high redshifts, this population can contain a great deal of the collapsed mass—from ∼10% at $z \sim 15$ to ∼30% at $z \sim 8$. Such objects are known as *minihalos*, and their large overdensities may have made them an important photon sink through the early stages of reionization.

However, these objects have shallow potential wells. As an ionization front reaches the halo, it heats the gas to $> 10^4$ K $> T_{\rm vir}$. Because the thermal pressure then exceeds the gravitational binding force, the minihalo gas escapes into the IGM through a strong evaporative wind. The high central densities of the minihalo gas transform the ionization front (which is R-type in the diffuse IGM) into a D-type front, generating a shock wave that expels the gas from the minihalo (see the discussion in §6.3).[35] This hydrodynamic process therefore occurs on roughly the sound-crossing time, $\sim c_s r_{\rm vir} \sim 30(M_h/10^7\ M_\odot)^{1/3}$ Myr at $z \sim 10$, which is much shorter than the age of the Universe at that redshift.

insert space / PE

One way to parameterize the effects of these minihalos on reionization is to supplement the IGM clumping factor with an average $C_{\rm mh} = \langle n_e^2 \rangle / \langle n_e \rangle^2$ over the minihalo density profiles. However, the rapid time evolution during evaporation makes application of this enhanced clumping factor difficult, because one must include each minihalo for only a finite time. A simpler parameterization is to use the total number of ionizing photons consumed (per minihalo atom) during the entire evaporation process. Detailed numerical simulations show that this process typically consumes ∼3–5 ionizing photons per

minihalo atom, as the high internal densities of the halos cause relatively rapid recombinations: $t_{\rm rec} \sim 2$ Myr (using the case-B rate) for a virialized object at $z \sim 10$.[36] Given the fraction of collapsed mass in these minihalos, this increases the number of photons per hydrogen atom required to complete reionization by about one, potentially making minihalos as important a photon sink as the clumped IGM itself. Fortunately, although these minihalos are clustered and so induce inhomogeneous recombinations, numerical simulations show that treating them as approximately uniform does not introduce any significant errors.[37]

Once a region is ionized, later formation of minihalos is strongly suppressed —even if the gas cools and recombines, because photoionization (or indeed any other substantial heating event, such as an X-ray background) dramatically increases the entropy of the IGM. In this context, the quantity

$$K = \frac{T}{n^{2/3}} = 760 \left(\frac{T}{10^4\ {\rm K}}\right) (1+\delta)^{-2/3} \left(\frac{1+z}{10}\right)^{-2}\ {\rm eV\ cm^2} \tag{9.64}$$

is usually referred to as "entropy," although the thermodynamic entropy is actually $S \propto \ln K$. Conveniently, K is conserved for any adiabatic process, including Hubble expansion or slow accretion; only strong shocks or radiative processes modify it. Clearly, the heating that occurs during reionization dramatically increases the entropy. Typical values are $K_{\rm reion} > 100$ eV cm^2 at $z \sim 10$, even after a substantial period of cooling and entropy release via recombination line cooling.

If this entropy is much larger than that generated by gravitational accretion onto a dark matter halo, the finite entropy "floor" prevents gas from collapsing to high densities—essentially preventing accretion onto the halo. It is convenient to parameterize this process in terms of the entropy generated by the accretion shock at the virial radius, which provides $K_{\rm halo} \approx T_{\rm vir}/[n(r_{\rm vir})]^{2/3}$. Interestingly, $K_{\rm reion}/K_{\rm halo} \sim 10(T_{\rm vir}/10^4\ {\rm K})^{-1}$ for an NFW profile; thus, the photoheating from reionization significantly suppresses accretion onto halos even somewhat above the usual atomic cooling threshold: numerical calculations of gas profiles (assuming hydrostatic equilibrium within the virial shock) show that only ∼50% of the gas is able to accrete when $K_{\rm reion}/K_{\rm halo} \sim 1$, and decreases rapidly for less massive halos.[38]

Photoheating suppresses accretion so efficiently because this process typically affects the gas while it has a low density and so efficiently imparts a large entropy to the gas. In fact, any other photoheating—even from a modest X-ray background generated by rare quasars—can substantially affect the IGM entropy, preventing the formation of minihalos even before they are ionized. We can use the estimate of equation (9.53) to examine this possibility as well: if a fraction f_h of the energy goes into heating (rather than ionization), we have

$$T_{\rm qso} \sim 20{,}000\, f_{\rm esc,q} \left(\frac{\epsilon}{0.1}\right) \left(\frac{f_{\rm UV}}{0.2}\right) \left(\frac{f_{\rm coll}}{0.01}\right) \left(\frac{f_{\rm BH}}{10^{-4}}\right) \left(\frac{f_h}{1/3}\right)\ {\rm K}, \tag{9.65}$$

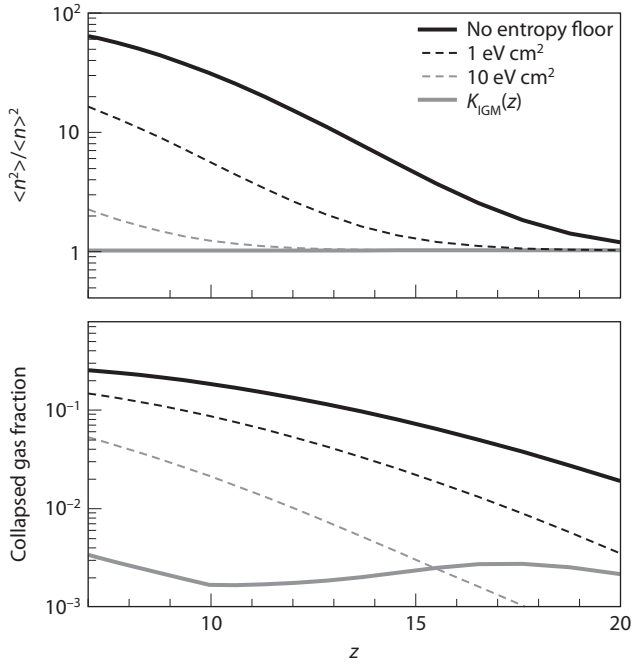


Figure 9.14 Effects of the IGM "entropy floor" on gas clumping from virialized minihalos (*top panel*) and the collapse fraction of gas onto dark matter halos (*bottom panel*). In each panel, the uppermost solid curve shows the model calculation with no entropy injection. The lower solid curve, labeled $K_{\rm IGM}(z)$, shows the effect of a numerical calculation of entropy injection via photoionization and subsequent recombination (which decreases K through radiative cooling). The two ~~dotted~~ curves show estimates for entropy injection at fixed levels (perhaps by an X-ray background). Oh, S. P., & Haiman, Z., *Mon. Not. R. Astron. Soc.* **346**, 456 (2005). Copyright 2005 by the Royal Astronomical Society.

dashed / AA

so substantial heating is clearly plausible. Even if $T_{\rm qso} \sim 1{,}000$ K—with a very modest accompanying ionized fraction—the arguments show that minihalo formation is almost completely suppressed.

Figure 9.14 shows some of these effects quantitatively. The bottom panel illustrates how the entropy suppresses the collapse of gas onto dark matter halos. The uppermost solid curve shows $f_{\rm coll}$ in this model if no excess entropy is introduced, and only minihalos with $T_{\rm vir} < 10^4$ K are included in the calculation. The ~~dotted~~ curves add $K = 1$ and 10 eV cm^2 (upper and lower, respectively). Even these modest levels reduce $f_{\rm coll}$ by a substantial amount. The lower solid curve, labeled $K_{\rm IGM}(z)$ shows a *minimal* suppression due to reionization, in which the gas is allowed to recombine for roughly a Hubble time (dramatically decreasing its entropy at high redshifts through recombination cooling). Even this conservative estimate essentially eliminates minihalo formation.

dashed / AA

The top panel shows an estimate of the effective clumping factor, $C = \langle n_e^2 \rangle / \langle n_e \rangle^2$, when only gas inside minihalos is included in the same scenarios as described previously. (Thus, it underestimates the *total* clumping factor, which must include gas outside virialized objects, but it more clearly shows the effect on these objects.) Again, even a relatively modest entropy injection dramatically reduces the role of these objects as photon sinks during reionization.

The suppression of accretion onto halos above the atomic cooling threshold is important for understanding high-redshift star formation. In detail this threshold depends on (*i*) self-shielding of gas within the potential well (which in turn depends on its internal structure); (*ii*) collisional recombination and cooling inside the halo; (*iii*) the amplitude of the ionizing background that impinges on each halo; and (*iv*) the relative timing of gas accretion onto the halo and the first appearance of the ionizing background.

Fortunately, simple arguments provide an estimate for the range of halos in which accretion is eventually suppressed. Halos larger than the Jeans mass in the heated medium are essentially unaffected; this is usually parameterized by a halo circular velocity threshold, V_J (see equation 3.31), with

$$V_{\rm J} = 81 \left(\frac{T_{\rm IGM}}{15{,}000 \text{ K}} \right)^{1/2} \text{ km s}^{-1}. \tag{9.66}$$

However, the dark matter halo itself actually has an average density ∼200 times the cosmic mean, so inside it the gravitational force gradient is larger than in the mean-density IGM (and hence better able to overcome thermal pressure). The Jeans mass evaluated with this larger density then determines the smallest halo that can accrete *any* gas, parameterized by the limiting circular velocity

$$V_{\rm lim} = 34 \left(\frac{T_{\rm IGM}}{15{,}000 \text{ K}} \right)^{1/2} \text{ km s}^{-1}. \tag{9.67}$$

Halos in the range from $V_{\rm lim}$ to $V_{\rm J}$ are able to accrete some, but not their entire complement of, gas. The point at which halos are able to accrete half the expected mass is roughly the filtering mass, or time-averaged Jeans mass (see §3.2). This is somewhat *smaller* than the Jeans mass itself because the thermal pressure is lower before reionization, which allows the early phases of assembly to proceed rapidly (and so to build up a halo near $V_{\rm lim}$).

The filter mass is a very useful approach for addressing this question, because (as a time integral) it illustrates how the feedback takes time to set in. Gas already close to accreting is still able to do so, because at the higher density characteristic of gas near halos, entropy injection is less efficient. This means that photoionization feedback manifests gradually over a timescale comparable to the collapse time of dark matter halos—essentially the Hubble time. Indeed, detailed simulations show that just after a given region is ionized, the suppression affects only halos with circular velocities $v_c < 10$ km s^{-1}.[39]

Because $V_{\rm J}$ typically lies above the atomic cooling threshold for star formation, reionization will *suppress* the formation of stars inside small galaxies. In principle, a search for such suppression provides another test of reionization models, although as described previously, suppression occurs gradually over a

timescale comparable to the Hubble time, so it will be difficult to separate from the many other factors that affect the cosmic star formation rate. If, however, reionization was highly inhomogeneous and extended over time, the differing reionization histories in different regions of the Universe may have induced variations in stellar populations whose observable effects persist to the present day. It may also have implications for understanding the wide range in stellar populations of Milky Way satellites with $V < V_J$, if some of those dark matter halos accreted gas (and formed their stars) before reionization, and some did so afterward. (see §13.4.2).

close up / PE

PART III
Observations of the Cosmic Dawn

Chapter Ten

Surveys of High-Redshift Galaxies

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. As discussed in the previous chapter, most theorists associate reionization with the first generations of stars, whose UV radiation streamed into intergalactic space and broke hydrogen atoms apart in H II bubbles that grew in size and eventually overlapped. Others conjecture that accretion of gas onto low-mass black holes gave off sufficient X-ray radiation to ionize the bulk of the IGM nearly simultaneously. New observational data are 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. The exploration of the reionization epoch promises to be one of the most active frontiers in cosmology over the coming decade. We are now in a position to understand the first pillar of these efforts: direct observations of galaxy populations.

What makes the study of the first galaxies so exciting is that it is a work in progress. Scientific knowledge often advances like a burning front, in which the flame is more exciting than the ashes. It will obviously be rewarding if our current theoretical ideas are confirmed by future observations, but it may be even more exciting if new observations demand that these ideas be modified.

10.1 Telescopes for Observing High-Redshift Galaxies

10.1.1 The Hubble Deep Field and Its Follow-Ups

In 1995, Bob Williams, then director of the Space Telescope Science Institute, invited leading astronomers to advise him where to point the Hubble Space Telescope (HST) during the discretionary time he received as a director, which amounted to a total of up to 10% of HST's observing time.[1] 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 toward a fixed, non-special direction and burn a hole in the sky as deep as we can go?"—much like

delete / PE

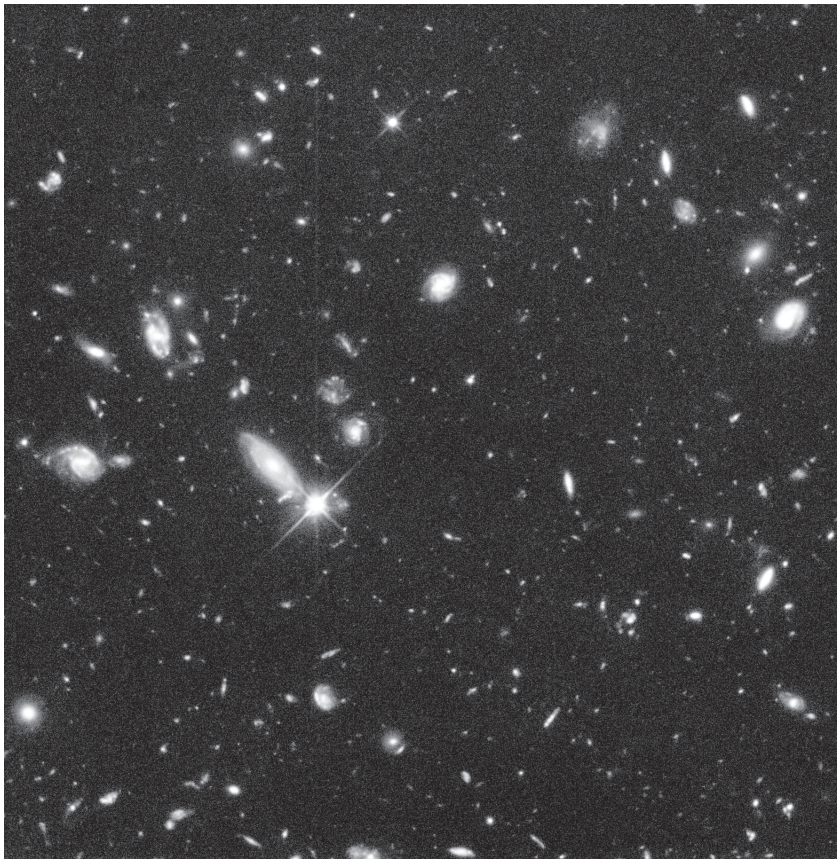


Figure 10.1 A portion of the first Hubble Deep Field (HDF) image taken in 1995 (see *Color Plate 18* for a color version of this image). The HDF covers an area 2.5 arcmin across and contains a few thousand galaxies (with a few candidates up to $z \sim 6$). The image was taken in four broadband filters centered on wavelengths of 3000, 4500, 6060, and 8140 Å, with an average exposure time of $\sim 1.27 \times 10^5$ s per filter.

testing how fast a new car can go. This simple compromise won the day, as 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; it produced the deepest image we have so far of the cosmos.

The Hubble Deep Field (HDF) covered an area of 5.3 arcmin^2 and was observed over 10 days (see Figure 10.1). One of its pioneering findings was the discovery of large numbers of high-redshift galaxies at a time when only a small number of galaxies at $z > 1$ were known: the HDF contained numerous red galaxies, with some reaching $z > 6$. The wealth of galaxies discovered at different stages of their evolutionary histories allowed astronomers to estimate the

variation in the global rate of star formation per volume over the lifetime of the universe.

Subsequent incarnations of this successful approach included the HDF-South (a near-replica of the original HDF in the southern sky) and the Great Observatories Origins Deep Survey (GOODS), which was a somewhat shallower survey covering a much larger area. Both of these extensions were designed to increase the original sample of galaxies in HDF, to make their statistics quantitatively reliable and to avoid spurious effects of clustering (see the detailed discussion in §10.4). A section of GOODS, occupying a tenth of the diameter of the full moon (equivalent to 11 arcmin^2), was then observed for a total exposure time of a million seconds to create the Hubble Ultra Deep Field (HUDF), the most sensitive (deepest) field image in visible light to date. Red galaxies have been identified in the HUDF image up to a redshift of $z \sim 8$, and possibly even higher, showing that the typical UV luminosity of galaxies declines with redshift at $z > 4$ (see §10.3). Most of the data we discuss in this chapter ultimately come from the HDF and HUDF.

10.1.2 Future Telescopes

The first stars emitted their radiation primarily in the UV band, but because of intergalactic absorption and their exceedingly high cosmological redshift, their detectable radiation is mostly observed in the IR (see §10.2.2). The successor to the Hubble Space Telescope, the James Webb Space Telescope (JWST), will include an aperture 6.5 m in diameter, made of gold-coated beryllium and designed to operate in the IR wavelength range of 0.6–28 μm (see Figure 10.2). JWST will be positioned at the Lagrange L2 point, where any free-floating test object stays on the side of earth away from the Sun but in a direct line with the Sun. JWST's large aperture and position outside Earth's atmosphere makes it particularly well suited to detecting the faint, compact galaxies we expect to have existed during the cosmic dawn and possibly to discover "smoking gun" signatures of Population III stars, such as strong UV sources with no metal lines or strong He II recombination lines (see §5.4).

capitalize / PE

Several initiatives to construct large IR telescopes on the ground are also underway. The next generation of ground-based telescopes will have effective diameters of 24–42 m, roughly three times wider than the largest existing optical/near-IR telescopes. Examples of these upcoming facilities include the European Extremely Large Telescope (EELT),[2] the Giant Magellan Telescope (GMT),[3] and the Thirty Meter Telescope (TMT),[4] which are illustrated in Figure 10.3. Along with JWST, they will be able to image and survey a large sample of early galaxies, and their large collecting areas will be especially useful in studying individual galaxies and their spectra in detail.

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 will potentially be detectable with the future Atacama Large Millimeter/Submillimeter Array (ALMA).[5] This array will contain sixty-six 7- to 12-m antennas positioned at very high altitudes in Chile, to see past the strong atmospheric

Figure 10.2 A full-scale model of the James Webb Space Telescope (JWST), the successor to the Hubble Space Telescope (http://www.jwst.nasa.gov/; see *Color Plate 19* for a color version of this image)). JWST includes a primary mirror 6.5 m in diameter and offers instrument sensitivity across the IR wavelength range of 0.6 to 28 μm, which will allow detection of the first generations of galaxies. The size of the Sun shield (the large flat screen in the image) is 22 m × 10 m (72 ft × 29 ft). The telescope will orbit 1.5 million km from Earth at the Lagrange L2 point. NASA/EPO.

delete second parenthesis / AA

Courtesy of /

absorption at millimeter and submillimeter wavelengths. It is perfectly positioned to observe emission from dust and heavy elements in the early Universe, as we discussed in §8.9.

Many other instruments are under development, to complement the direct views of the galaxies obtainable with these telescopes. For example, given that these galaxies also created ionized bubbles during reionization, their locations should be correlated with the existence of cavities in the distribution of neutral hydrogen. Within the next decade it may become feasible to explore the environmental influence of galaxies by using IR telescopes in concert with radio observatories that will map diffuse hydrogen at the same redshifts (see §12 and §13.3).

10.2 Methods for Identifying High-Redshift Galaxies

Much of the baryonic mass in the Universe assembled into star-forming galaxies after the first billion years in cosmic history. Consequently, the highest-redshift galaxies are a rarity among all faint galaxies on the sky. A method for isolating candidate high-redshift galaxies from the foreground population of feeble lower-redshift galaxies is required to identify targets for follow-up studies.

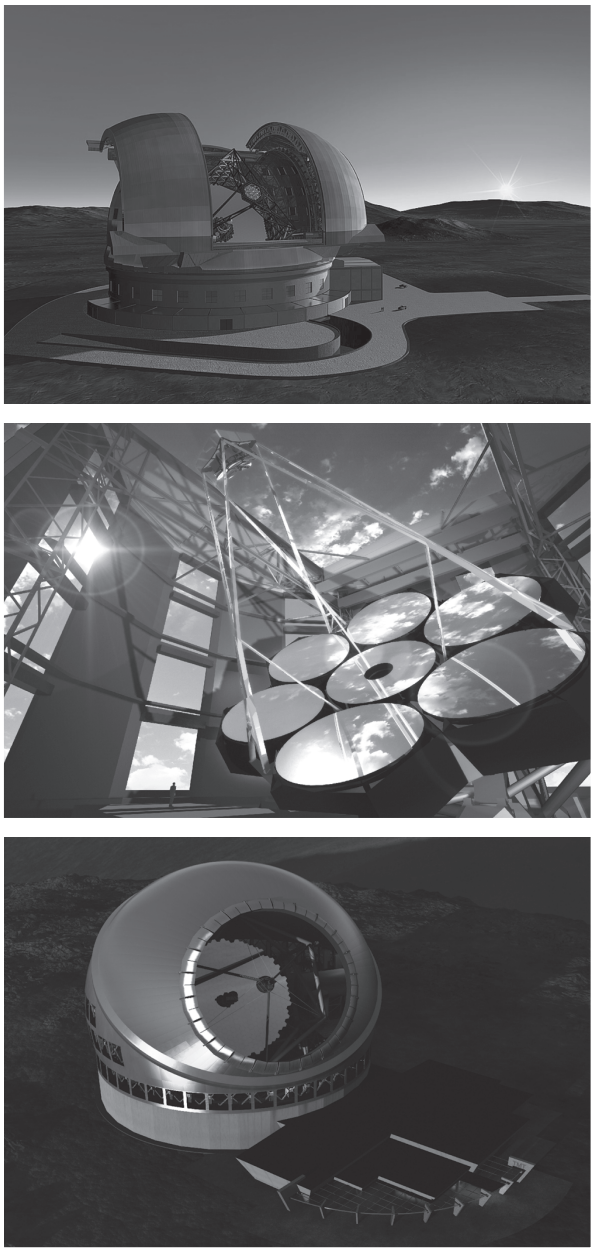


Figure 10.3 Artist's conception of the designs for three future giant ground-based telescopes that will be able to probe the first generation of galaxies (see *Color Plate 20* for a color version of this image): the European Extremely Large Telescope (EELT, top), the Giant Magellan Telescope (GMT, middle), and the Thirty Meter Telescope (TMT, bottom). The European Southern Observatory (ESO), the GMT Partnership, and the TMT Observatory Corporation.

Karen: The last sentence lists the sources. Add "Courtesy of" to beginning?

10.2.1 Lyman-α Emitters

PE

One technique makes use of narrow-band imaging to identify galaxies whose highly redshifted line emission falls within the selected band. An object that is bright in the narrow band but faint (or, for these applications, usually invisible) in nearby broadband measurements can be identified as a line emitter. Provided one can identify the line that is observed, this technique has the advantage of identifying both the redshift and location of the galaxy. This method is typically applied to the Lyman-α line, which is often very strong because most ionizing photons absorbed by the galaxy's ISM are reprocessed into Lyman-α line photons through recombinations (see chapter 11). However, it is also highly sensitive to the gas geometry and kinematics and can be extinguished by dust. The galaxies detected by this technique are termed *Lyman-α emitters (LAEs)*. The following are the primary challenges with this approach:

- *The infrared night sky:* Terrestrial telescopes suffer from substantial atmospheric absorption and strong night-sky lines in the IR bands (primarily from OH and water vapor). Figure 10.4 shows the night sky in the relevant spectral range, including both atmospheric absorption and night-sky emission lines. The vertical shaded columns show "windows" where the emission lines are below one-third of the average. The dark and the light regions take moderate- and high-resolution bands, respectively (with $R = \lambda/\Delta\lambda = 300$ and 1,000). These open bands cover only 16% and 27% of the available spectrum, respectively, indicating that this technique can be used only in particular redshift ranges. So far, the most commonly utilized are at $z \sim 6.6$, 7, 7.7, and 8.5.
- *Contamination from lower-z line emitters:* Galaxies have many other emission lines, of course, some of which can be very strong. Of particular concern are Hα, [O III], Hβ, and [O II] lines. With only a single line detection, a firm identification that distinguishes among these possibilities cannot be made. Such contaminants must be ruled out by detecting other emission lines from the source (unlikely to be visible for a true LAE but very plausible for the lower-redshift interlopers) or by measuring the continuum emission (obviously very difficult for a faint source). If only a single line is visible, the shape can help determine whether the object is truly an LAE: as shown in Figure 10.5, observed Lyman-α lines in galaxies nearly always have asymmetric profiles, with a sharp cutoff on the blue side (due to IGM absorption) and a long tail to the red side (due to radiative transfer effects). Metal lines, in contrast, are generally very symmetric. Unfortunately, while very suggestive, these types of line shape diagnostics are not perfect. A robust identification always relies on multiple lines. The next best option is deep follow-up of the source to observe the continuum break described in the next section.
- *Interpretation and follow-up:* Finally, although this method efficiently finds galaxies at high redshifts, it provides little direct physical information—only a single-line luminosity that, as we will see in chapter 11, is

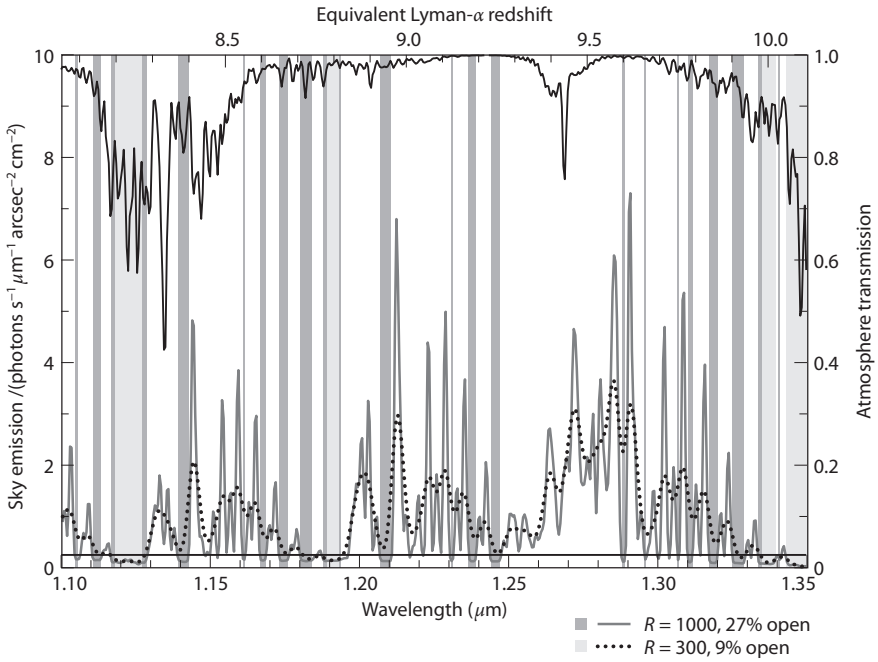


Figure 10.4 "Windows" in the J-band night-sky spectrum. The black line indicates the transmission of the night sky (scale on right). The two lines at the bottom show the night-sky spectrum at two resolutions ($R = 1000$ and 300; intensity scale on left). The vertical shading shows regions where the emission is less than one-third of the mean value. Barton, E. J., et al., *Astrophys. J.* **604**, L1 (2004). Reproduced with permission of the American Astronomical Society.

heavily dependent on dust, the ISM clumpiness and dynamics, and the IGM ionization state. Even deep follow-up observations typically detect little or no stellar continuum emission.

To date, LAE surveys have detected many high-z sources, but their interpretation is still debated. We return to the Lyman-α line as an important cosmological probe in chapter 11.

10.2.2 Lyman-Break Galaxies

The second observational technique adopts several broad photometric bands to estimate the redshifts of galaxies based on the strong spectral break arising from absorption by intergalactic (or galactic) neutral hydrogen along the line of sight to the source. As we saw in chapter 4, the IGM is optically thick to Lyman-α photons at high redshifts. Thus, little or no flux should be detectable shortward of 1216 $(1 + z)$ Å (irrespective of the history of reionization). For example, to identify a galaxy at $z = 6$ one needs two filters: one above and

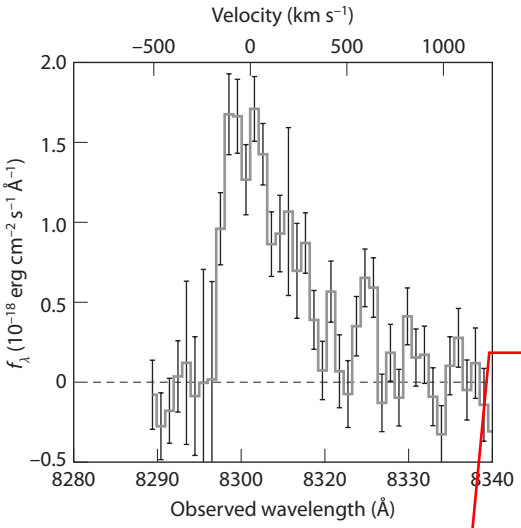


Fig. 6B. /

Figure 10.5 The Lyman-α emission line of a typical i'-dropout galaxy SBM03# at $z = 5.83$. Stanway, E., et al., *Astrophys. J.* **607**, 704 (2004). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /

AAS. /

the other below the Lyman-α break at $7 \times 1216 = 8512$ Å. The relevant bands are i' (centered at ~9000 Å) and z' (centered at ~8000 Å) of HST, as illustrated in Figure 10.6. This method was first used at lower redshifts, $z \sim 3$–4, where the intergalactic H I column density is smaller, and so the related Lyman-limit break at 912 Å was instead adopted to photometrically identify galaxies. The 912 Å 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 technique are termed *Lyman-break galaxies (LBGs)*.

insert spaces / PE

The key challenge for observers is to obtain a sufficiently high signal-to-noise ratio that LBGs can be safely identified through the detection of a single redder band. Figure 10.6 illustrates how a color cut of $(i' - z')_{AB} > 2.3$ (see equation 1.16 for a definition of the AB magnitude system) is effective at selecting sources at redshifts $z > 6$. The reliability of this dropout technique in rejecting low-redshift interlopers can be tested only through spectroscopic observations. The i'-drop spectra typically show a single emission line at the Lyman-α wavelength, with no significant continuum; as in Figure 10.5, the lines are typically asymmetric and can clearly indicate the source redshift. However, only a fraction of galaxies have Lyman-α lines, and spectroscopic follow-up is often difficult.

The NIRSpec spectrograph on JWST covers observed wavelengths in the range 0.8–5 μm and is ideally suited for the task of identifying the redshifts of distant galaxies. This instrument has the sensitivity to detect the rest-frame UV and optical continuum emission over the full range of emission lines from Lyman-α to Hα (6563 Å rest wavelength) for galaxies at $z \sim 6$. Analogous

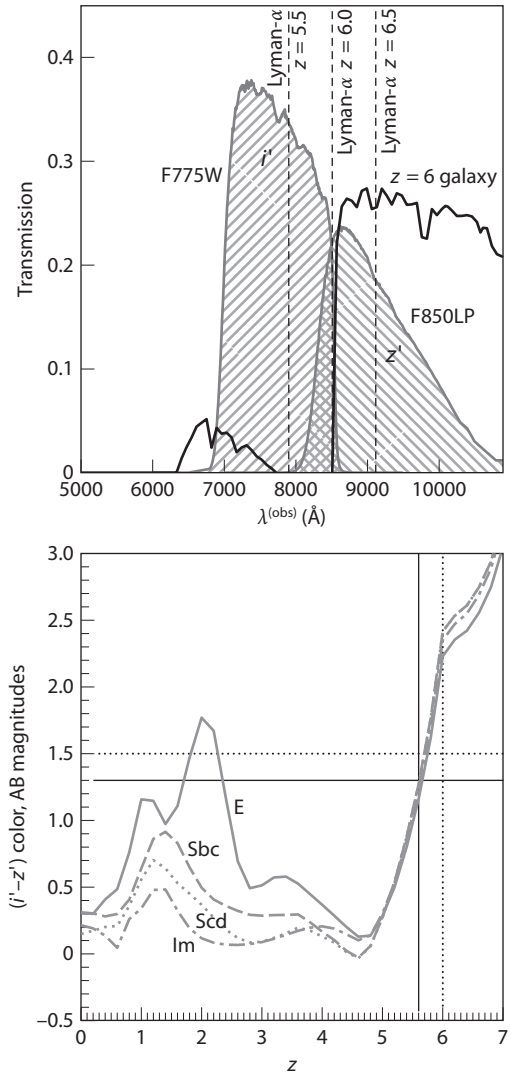


Figure 10.6 *Top panel:* The i' and z' bands of HST (shaded regions) on top of the generic spectrum from a galaxy at a redshift $z = 6$ (solid line). The Lyman-α wavelength at various redshifts is also shown. *Bottom panel:* Models of the color-redshift tracks for different types of galaxies with nonevolving stellar populations. All the standard galaxy types converge to have very red $(i' - z')$ colors at high redshifts, so such galaxies can be efficiently identified with a cut in color space. The bump at $z \sim 1$–2 arises when the Balmer break or the 4000 Å break redshift beyond the i'-filter. Synthetic models indicate that the Balmer break takes $\sim 10^8$ years to establish, providing a measure of the galaxy age. Bunker, A., et al. *Frontiers of Astrophysics: A Celebration of NRAO's 50th Anniversary*, ASP Conference Series, vol. 395, ed. A. Bridle, J. Condon, & G. Hunt (2008). Copyright 2008 by the Astronomical Society of the Pacific.

studies of galaxies at $z \sim 3$ with HST have produced a detailed understanding of the internal properties of these galaxies.

10.2.3 Using Massive Gravitational Lenses as Natural Telescopes

A massive gravitational lens, such as a cluster of galaxies, can be used to probe deeper into the early Universe and detect faint sources that are otherwise below the sensitivity of human-made telescopes (see §8.10 for an overview of the physics of this process). Foreground X-ray clusters can provide a magnification boost of 5–30 in flux (for unresolved sources) or in size (for resolved sources). The gain in reaching a fainter flux threshold is partly offset by a reduction in the sky area being surveyed behind the lens, which produces an overall change in the number of bright sources that depends on the slope of the luminosity function (see equation 10.10).

The inverse of the 2×2 magnification matrix, defined as $\mathcal{M}^{-1} \equiv \partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}$, is real and symmetric and can therefore be diagonalized and expressed along its principal axis as

$$\mathcal{M}^{-1} = \begin{pmatrix} 1-\kappa-\gamma & 0 \\ 0 & 1-\kappa+\gamma \end{pmatrix}. \tag{10.1}$$

The so-called *convergence coefficient* κ is associated with an isotropic focusing of light rays (and is given by the surface density of the lens scaled by $\Sigma_{\rm crit}$), while the *shear coefficient* γ (defined by the trace-free component of the matrix and derived from the lens equation) introduces anisotropy. A circular source of unit radius is distorted to an elliptical image with major and minor axes of $(1-\kappa-\gamma)^{-1}$ and $(1-\kappa+\gamma)^{-1}$, respectively. Since lensing conserves surface brightness, the magnification μ is the ratio between the image area and the area of the source or, equivalently, the determinant of the magnification matrix,

$$\mu = \det \mathcal{M} = \frac{1}{\det(\mathcal{M}^{-1})} = \frac{1}{(1-\kappa)^2-\gamma^2}. \tag{10.2}$$

For a point source, the magnification diverges at the so-called critical lines in the image plane of the lensing cluster (or the equivalent "caustics" in the source plane). The magnification is infinite if one of the principal values of $\mathcal{M}^{-1}$ equals zero. In the image plane, this condition defines two closed lines that do not intersect. Finite-size sources, like galaxies, are magnified only by a finite factor, because most of their observed light rays cannot get closer to a caustic than the source size. For simple mass distributions, there are two critical lines: the external critical line where the source is deformed in the *tangential* direction, and the internal critical line where the deformation is *radial*.

For a circularly symmetric surface-density profile (as in the case of a spherically symmetric lens), the critical lines are circles. The tangential critical line in this case is located at the Einstein radius, interior to which the *average* surface density equals the critical value $\Sigma = \Sigma_{\rm crit} = 0.35\ {\rm g\ cm^{-2}}(D/1\ {\rm Gpc})^{-1}$. This provides a simple method for measuring the total projected mass of the lens out to that radius if the source and lens redshifts are known. The location of the

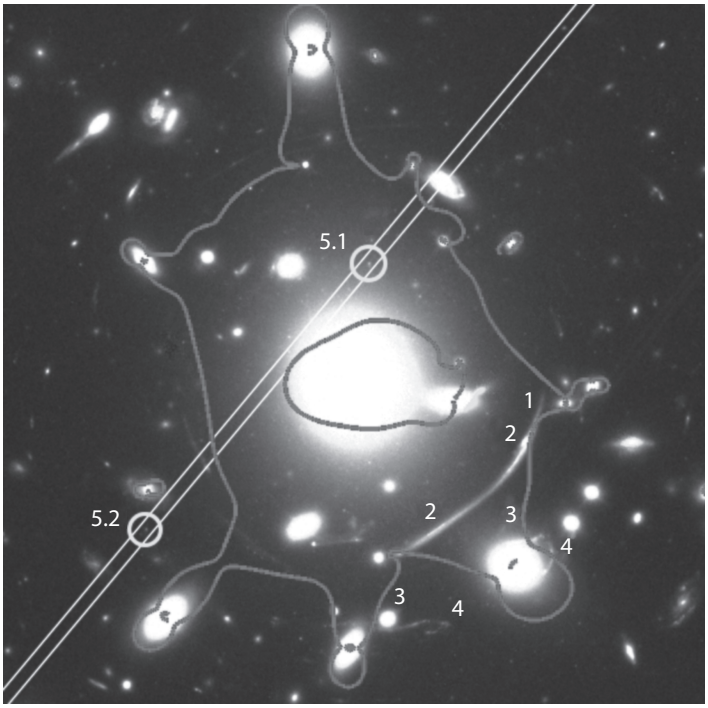


Figure 10.7 Image of the core of the cluster Abell 383, including known multiply imaged sources (marked 1 to 5) and the noncircular critical lines for sources at $z_s = 6$ (see *Color Plate 21* for a color version of this image). The two images of a galaxy with $z_s = 6.027$ are marked by circles. The long slit used for spectroscopic follow-up is shown in white. Richard, J., et al., *Mon. Not. R. Astron. Soc.* **414**, L31 (2011). Copyright 2011 by the Royal Astronomical Society.

radial critical line depends on the inner gradient of the mass distribution. Deviations from circular symmetry complicate the critical line geometry, which in the general case needs to be solved numerically. Clearly, the location and shape of the critical lines depend on the source redshift.

Figure 10.7 shows an example for lensing of a galaxy at a redshift $z_s = 6.027$ by the cluster Abell 383 at $z_d = 0.187$. The inner and outer critical lines have a noncircular geometry owing to the ellipticity and substructure in the cluster mass distribution. The two images of the background galaxy, labeled 5.1 and 5.2, are marked by circles. The lens model implies magnification factors of $\mu_1 = 11.4 \pm 1.9$ and $\mu_2 = 7.3 \pm 1.2$ for images 5.1 and 5.2, respectively. These factors give an unlensed AB magnitude of 27.2 ± 0.05 for the source galaxy in the H band. At $z = 6$, 1" on the sky corresponds to a projected distance of 5.7 kpc.

Recent estimates of the faint-end slope of the luminosity function of star-forming galaxies at $z > 6$ suggest that the bulk of the integrated star

formation at high redshift may originate in low-luminosity galaxies, as expected theoretically. To clarify whether low-luminosity systems are abundant at early times requires observations probing well below the sensitivity limits obtainable with current facilities using conventional methods. Strong gravitational lensing by foreground clusters is the only means to advance this quest before the next generation of telescopes is available. Given the finite size of background galaxies, a typical lensing cluster can magnify faint galaxies at $z > 7$ by more than a factor of $\sim$15 on the critical lines. Spectroscopy and detailed studies become feasible at otherwise impossible unlensed limits. This technique enables the determination of stellar UV continuum slopes, Lyman-α emission line profiles, and star formation rates for sources whose intrinsic flux is close to the faintest limits reached with direct imaging (AB magnitude $\sim$30 in the I-band).

Surveys for gravitationally lensed high-redshift galaxies have been conducted using two techniques. The first technique makes use of long-slit spectrographs on 8- to 10-m ground-based telescopes to "scan" the cluster critical line to search for high-redshift strongly lensed Lyman-α emitting galaxies. By focusing on the areas of highest magnification, this technique is sensitive to the lowest-luminosity galaxies at high redshifts. The critical-line mapping technique was extended into the near-IR to search for lensed Lyman-α emitters in the redshift interval $8.5 < z < 10.2$, where candidate sources were discovered.[6]

The second technique makes use of sensitive multiwavelength imaging of galaxy clusters to select galaxies using the standard Lyman-break technique. This method generally covers a much larger area than the critical-line survey, albeit with significantly lower average magnifications. Early studies of lensed dropouts with HST have demonstrated the potential of this method for locating $z > 6$ galaxies. Deep near-IR imaging identified a large-volume density of faint dropouts likely to lie at $z > 6$.[7] As with the critical-line surveys, these results hinted at a large ionizing contribution from low-luminosity galaxies.

Lensing surveys will improve as more efficient instrumentation is installed on existing facilities. One of the most exciting developments has been the installation of the Wide Field Camera 3 (WFC3) onboard HST, which is providing deeper IR imaging over a larger area than previously available. WFC3 imaging of the clusters Abell 1703 and Abell 383 have yielded a number of promising $z > 6$ galaxies, as illustrated in Figure 10.7. The magnification provided by gravitational lensing enables detection of galaxies with a high signal-to-noise ratio and results in more secure measurements of the UV continuum slope and stellar mass. Moreover, the magnified galaxies are ideal for spectroscopic follow-up. Follow-up observations of one of the lensed galaxies in Abell 1703 produced the first robust spectroscopic confirmation of a $z > 7$ Lyman-break galaxy using a near-IR spectrograph.[8] The prevalence of Lyman-α emission in these galaxies is being used to set constraints on when reionization occurred. Deeper follow-up spectroscopy of these galaxies will provide strong constraints on the strength of He II λ1640 emission and offer the potential to identify the telltale UV signatures of hot Population III stars.

10.2.4 Finding Faint Galaxies with the First Gamma-Ray Bursts

Traditional methods of finding galaxies, including both the LAE and LBG techniques, select galaxies above a given luminosity threshold and so are biased toward identifying the brightest galaxies. However, as we have seen, much of the activity at high redshifts likely occurs in faint galaxies far below the luminosity threshold of even extremely deep observations like the HUDF. *Is there any way to find more typical galaxies?*

Remarkably, the best way may be to search for individual "stars" rather than the collective emission of entire galaxies. Explosions of individual massive stars (such as supernovae) can outshine their host galaxies for brief periods of time. The brightest among these explosions are gamma-ray bursts (GRBs), observed as short flashes of high-energy photons followed by afterglows at lower photon energies (as discussed in §5.6). These afterglows can be used to study the first stars directly. To date, GRBs have been discovered by the *Swift* satellite out to $z = 9.4$, merely 540 million years after the Big Bang, and significantly earlier than the farthest known quasar ($z = 7.1$).[9] It is already evident that GRB observations hold the promise of opening a new window into the infant Universe.

As discussed in §5.6, long-duration GRBs are believed to originate from the collapse of massive stars at the end of their life (Figure 5.18). Since the very first stars were likely massive, they could have produced GRBs. If so, we may be able to see their host galaxies one star at a time. The discovery of a GRB afterglow whose spectroscopy indicates a metal-poor gaseous environment could potentially signal the first detection of a Population III star. The GRB redshift can be identified from the Lyman-α break in its otherwise power-law UV spectrum. A photometric detection can then be followed up with spectroscopy on a large telescope. Various space missions are currently proposed to discover GRB candidates at the highest possible redshifts.

In addition to being valuable in individual source detections, GRBs are expected to reside in typical small galaxies where massive stars form at those high redshifts. Once the transient GRB afterglow fades away, observers may search for the steady but weaker emission from its host galaxy. High-redshift GRBs may therefore serve as signposts of high-redshift galaxies that are otherwise too faint to be identified on their own. Importantly, GRBs trace the star formation history in a different way than do typical galaxy surveys, because they can reside in arbitrarily faint galaxies below any realistic survey detection threshold (although other biases, such as metallicity, may be important).

Although standard lightbulbs appear fainter with increasing redshift, this is not the case with GRBs, because they are transient events that fade with time. When studying a burst at a constant *observed* time delay, we are able to see the source at an earlier time in its own frame. This is a simple consequence of time stretching due to the cosmological redshift. Since the bursts are brighter at earlier times, it turns out that detecting them at high redshifts is almost as feasible as finding them at low redshifts, when they are closer to us.[10] It is a fortunate coincidence that the brightening associated with seeing the GRB at an intrinsically earlier time roughly compensates for the dimming

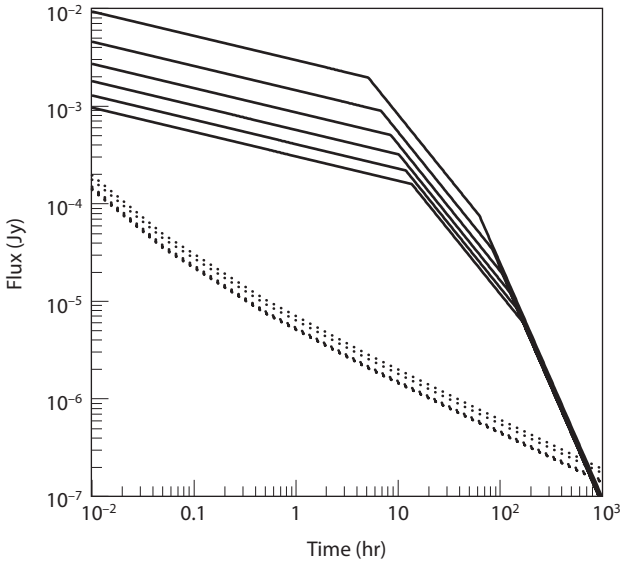


Figure 10.8 Detectability of high-redshift GRB afterglows as a function of time since the GRB explosion as measured by the observer. The GRB afterglow flux (in Jy) is shown at the redshifted Lyman-α wavelength (solid curves). Also shown (dotted curves) is a crude estimate for the spectroscopic detection threshold of the 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). Barkana, R., & Loeb, A., *Astrophys. J.* **601**, 64 (2004). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

associated with the increase in distance to the higher redshift, as illustrated by Figure 10.8.

10.3 Luminosity and Mass Functions

The luminosity function (LF) of galaxies, $\phi(L)\,dL$, describes the number of galaxies per comoving volume within the luminosity bin between L and $L + dL$. It is the most fundamental observable quantity for galaxy surveys, and a great deal of effort has gone into measuring it in both the nearby and distant Universe. Figure 10.9 shows measurements at $z = 4$–8 of the rest-frame UV galaxy luminosity function, with the most distant data taken from the HUDF.

A popular fitting form for a wide range of galaxy surveys is provided by the *Schechter function*,

$$\phi(L) = \phi_\star \left(\frac{L}{L_\star}\right)^{\alpha} \exp\left(-\frac{L}{L_\star}\right), \tag{10.3}$$

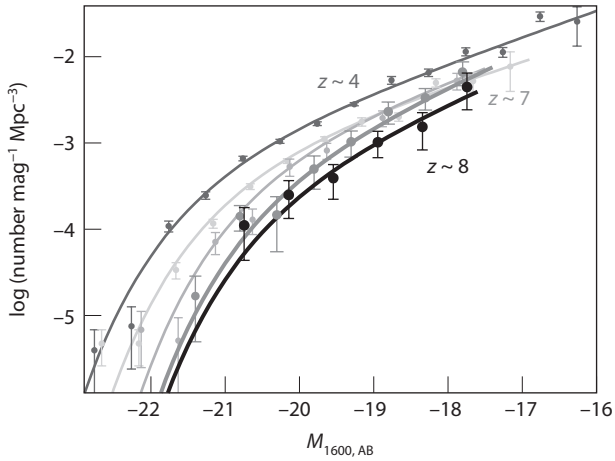


Figure 10.9 Rest-frame UV luminosity functions derived for galaxies at $z \sim 7$ (large filled circles) and $z \sim 8$ (large open circles) compared with lower-redshift data (from $z = 6$ to $z = 4$; the density increases as redshift decreases); see *Color Plate 22* for a color version of this image. The vertical axis gives the number of galaxies per comoving cubic megaparsecs per AB magnitude at a rest-frame wavelength of 1600 Å, as a function of this magnitude on the horizontal axis. Note the sharp decline in the number density of bright galaxies with redshift and tentative evidence for a steepening faint-end slope. Bouwens, R., et al., *Astrophys. J.* **737**, 90 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

where the normalization $\phi_\star$ corresponds to the volume density at the characteristic luminosity $L_\star$ and α is the faint-end slope that controls the relative abundance of faint and bright ($L > L_\star$) galaxies. The total number density of galaxies is given by $n_{\rm gal} = \int_0^\infty \phi(L)\,dL = \phi_\star \Gamma_G(\alpha+1)$, and the total luminosity density is $u_{\rm gal} = \int_0^\infty \phi(L) L\,dL = \phi_\star L_\star \Gamma_G(\alpha+2)$, where Γ_G is the Gamma function. Note that at the faint end, $n_{\rm gal}$ diverges if $\alpha < -1$, and $u_{\rm gal}$ diverges if $\alpha < -2$. (In reality, the integrals converge anyway because there is a minimum luminosity for galaxies, set by a combination of the minimum halo mass for gas accretion and the minimum halo mass in which gas can cool.)

The curves in Figure 10.9 show fits of this form to the data; clearly, this simple empirical structure does an excellent job of matching the observations. Three points emerge from these fits: *(i)* the characteristic luminosity $L_\star$ declines toward higher redshift; *(ii)* the space density of galaxies at $L_\star$ also decreases; and *(iii)* the faint-end slope may steepen at $z > 7$ (though the evidence for this is still tentative). In light of typical models for structure formation, in which these galaxies are associated with dark matter halos, these results are hardly surprising: at higher redshifts, fewer halos have formed, and in any hierarchical model those that have formed are preferentially smaller. The interesting physics involves the mapping from the halo mass function to luminous baryons, which we discussed in chapter 8.

The particular physical insight provided by a galaxy survey depends on the selection technique and wave band used. In general, rest-frame UV measurements (such as those shown in Figure 10.9) depend exclusively on hot stars able to produce the observed UV photons. Because these high-mass stars are short-lived, the UV luminosity is tied to the star formation rate (SFR) of the galaxies, although there is an uncertain correction that depends on the stellar IMF, because the ratio of high-mass to low-mass stars determines the efficiency with which baryons produce high-energy photons. In other words, the high-energy photons provide little direct information about low-mass stars, so the total SFR requires some extrapolation. In fact, there are several ways to estimate SFRs from astronomical measurements.[i]

- *The rest-frame UV continuum* (1250–1500 Å) provides a direct measure of the abundance of high-mass $>5\ M_\odot$ main-sequence stars. Since these stars are short-lived, with a typical lifetime $\sim 2 \times 10^8 (m_\star/5\ M_\odot)^{-2.5}$ yr, they provide a good measure of the "instantaneous" SFR, with

$$\mathrm{SFR} \approx 1.4 \left(\frac{L_\nu}{10^{28}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}} \right) M_\odot\ \mathrm{yr}^{-1}. \tag{10.4}$$

  The primary uncertainty in this determination is extinction via dust, though that can be estimated from the spectra or from other observations.
- *Nebular emission lines*, such as Hα and [O II], measure the combined luminosity of gas clouds that are photoionized by very massive stars ($>10\ M_\odot$). The total amount of line emission therefore measures the rate at which ionizing photons are produced and hence the (massive) SFR. 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,

$$\mathrm{SFR} \approx 0.8 \left(\frac{L(\mathrm{H}\alpha)}{10^{41}\ \mathrm{erg\ s^{-1}}} \right) M_\odot\ \mathrm{yr}^{-1}, \tag{10.5}$$

  and

$$\mathrm{SFR} \approx 1.4 \left(\frac{L([\mathrm{O\ II}])}{10^{41}\ \mathrm{erg\ s^{-1}}} \right) M_\odot\ \mathrm{yr}^{-1}. \tag{10.6}$$

- *Far-IR emission* (10–300 μm) measures the total emission from dust heated by young stars. Again, this heating is due to high-energy radiation from massive stars, and so the total luminosity measures how rapidly they form:

$$\mathrm{SFR} \approx 0.45 \left(\frac{L(\mathrm{FIR})}{10^{43}\ \mathrm{erg\ s^{-1}}} \right) M_\odot\ \mathrm{yr}^{-1}. \tag{10.7}$$

- *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

[i] Unless otherwise specified, the following conversions assume a standard Salpeter IMF with solar metallicity to transform luminosity into stellar mass.

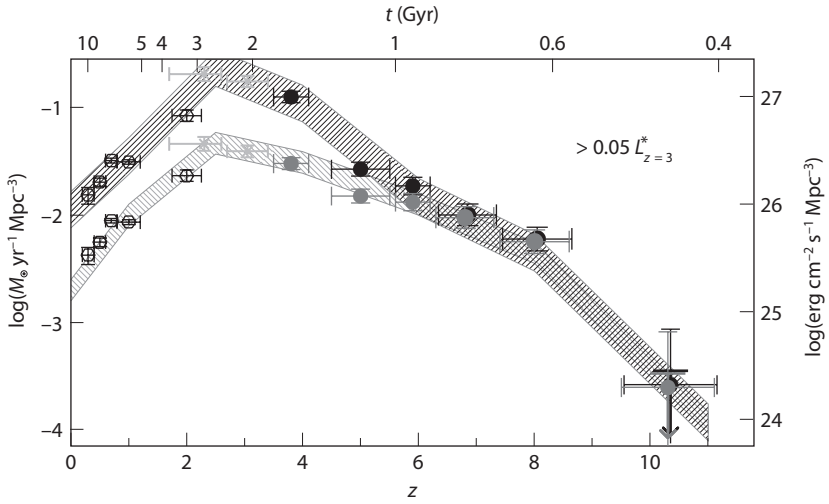


Figure 10.10 The star formation rate density (left vertical axis) and luminosity density (right vertical axis) as functions of redshift (lower horizontal axis) and cosmic time (upper horizontal axis), for galaxies brighter than an AB magnitude of –17.7 (corresponding to $0.05L_\star$ at $z = 3$). The conversion from observed UV luminosity to star formation rate assumed a Salpeter IMF for the stars. The upper curve includes dust correction based on estimated spectral slopes of the observed UV continuum. Bouwens, R., et al., *Astrophys. J.* **737**, 90 (2011). Reproduced with permission of the American Astronomical Society.

of massive stars ($>8\,M_\odot$), because these have a short lifetime, giving on timescales longer than $\sim 10^8$ yr,

$$\mathrm{SFR} \approx 1.1 \left(\frac{L_\nu(1.4\,\mathrm{GHz})}{10^{28}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}} \right) M_\odot\ \mathrm{yr}^{-1}. \qquad (10.8)$$

We emphasize again that all these measures are calibrated locally, and their extrapolation to high redshifts is at best uncertain.

Integration of the luminosity function of galaxies over a kernel that measures their SFR rate yields the SFR per comoving volume in the Universe. Figure 10.10 shows a recent determination of this rate as a function of redshift based on measurements of the UV luminosity function of all galaxies brighter than $0.05L_\star$ at $z = 3$ (corresponding to an AB magnitude of –17.7). Most of these measurements are made from UV data, so the correction for dust extinction is particularly important (the corrected form is shown by the upper set of measurements here). Also, we note that this is a *lower* limit to the true SFR density, because it ignores feeble galaxies below the detection threshold.

One obvious omission from our list of SFR indicators is the Lyman-α line, which, as we discussed, is useful in detecting high-redshift galaxies. However, as we will see in chapter 11, the interpretation of this emission line is fraught

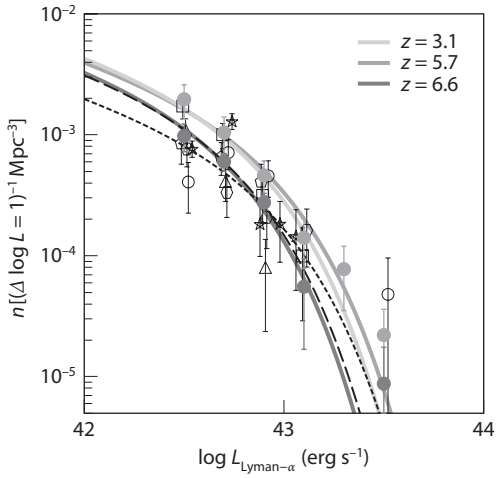


Figure 10.11 Luminosity function of LAEs for the range $z = 3$–6.6 (see *Color Plate 23* for a color version of this image). The light solid circles show the measured luminosity function at $z = 6.6$, while the darker solid circles show the same for $z = 5.7$. The solid lines show Schechter function fits to these as well as the best fit at $z = 3.1$ (lightest curve). The LAE density drops substantially from $z = 5.7$ to $z = 6.6$, much faster than that of LBGs, but is nearly constant at lower redshifts. Finally, the open symbols show the number densities measured in the five subfields of the $z = 6.6$ survey, illustrating the substantial variance between fields. Ouchi, M., et al., *Astrophys. J.* **723**, 869 (2010). Reproduced with permission of the American Astronomical Society.

by / AAS. /

with uncertainties about the galaxy's dust content, ISM structure, outflow properties, and environment. Therefore, the Lyman-α line is not a good SFR indicator. However, one can still construct a luminosity function of emission in this line; Figure 10.11 shows recent determinations in the redshift range of $z = 3$–6.6. In contrast with the luminosity function of photometrically selected LBGs, LAEs do not appear to change in comoving number density in the range $z = 3$–5.7, although their density appears to decline rapidly beyond that. This may be an indication of changes in the galaxy environments—and possibly reionization—though we will describe many difficulties with such an interpretation in chapter 11.

Meanwhile, the mass budget of stars at $z \sim 5$–6 can be inferred from their rest-frame optical and near-IR luminosities, which are much closer to measuring the total stellar content than UV light, because even low-mass stars emit in these bands. In many ways, the total stellar content is more physically interesting than the SFR density, because the cumulative density provides a census of stars that previously formed inside faint galaxies below the detection threshold and only later were incorporated into detectable galaxies. Moreover, the cumulative density provides some information on the *past* history of star formation

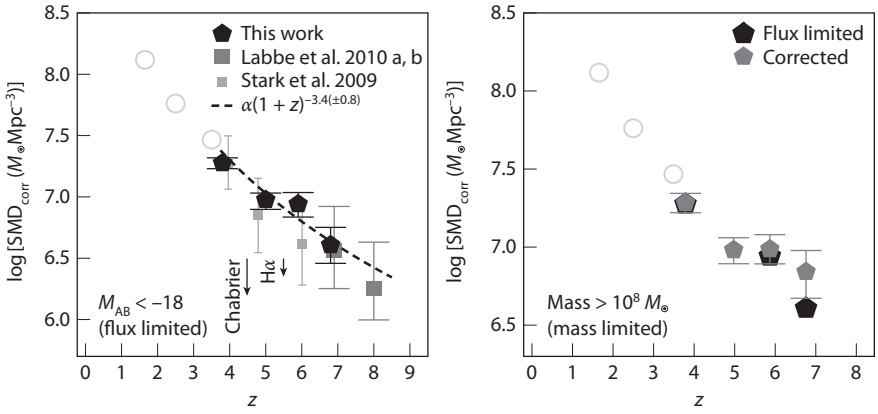


Figure 10.12 Stellar mass density (SMD) evolution over redshift in galaxies brighter than $M_{AB} = -18$ at a rest-frame wavelength of 1500 Å (*left*) and with stellar mass $M_\star > 10^8\ M_\odot$ (*right*) from several different measurements. All panels assume a Salpeter IMF and $Z = 0.2\ Z_\odot$; the line labeled "Chabrier" shows the effect of assuming a different IMF. The other line (labeled "Hα") shows the effect of systematic contamination from line emission. Gonzalez, V., et al., *Astrophys. J.* **735**, L34 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

(though still subject to uncertainty with the IMF). Figure 10.12 shows some recent measurements of the growth of the total stellar mass density in the Universe. Note in particular that only a small fraction of the stars present at $z < 2$ formed at $z > 6$, though this finding is subject to an unknown correction from undetected galaxies.

10.3.1 Emission Lines

Because the Lyman-α line can be very bright, and because its UV rest wavelength redshifts it into the optical or near-IR in distant galaxies, this line gets a great deal of attention (we discuss it extensively in chapter 11). But other emission lines can be as or even more useful for certain diagnostics, and we briefly mention them here. This is of course a very extensive field of research, and so we refer the interested reader to the literature and other textbooks for more information (see Further Reading).

1. *Other hydrogen lines:* The other Lyman-series lines are almost never visible in high-redshift galaxies; after several scatterings, these photons are "recycled" via radiative cascades into either Lyman-α photons or a pair of photons from the forbidden $2s \rightarrow 1s$ decay (see §12.2.2). However, Balmer-series photons (and those beginning at even higher levels) are very useful diagnostics. They are initially generated through the same process as Lyman-α—recombinations following ionizations near hot, massive stars—but because such photons can interact only with atoms already in

the $n = 2$ state, they are not subject to scattering in the ISM and escape galaxies relatively easily (especially since they have relatively long wavelengths and so are less subject to dust absorption; e.g., the Hα line lies at 6563 Å). They therefore offer much more robust measures of star formation rates, subject only to the uncertainty in the IMF and dust (see equation 10.5).

Unfortunately, although Hα is extremely important for low-redshift galaxies, its relatively long rest wavelength has so far limited its utility for high-redshift galaxies.

2. *Helium lines:* He II has the same electronic structure as H I, but shifted to four times larger energies. As a result, its ionization potential is well beyond the cutoff of most stars—only rare Wolf-Rayet stars (i.e., massive stars undergoing rapid mass loss) and the most massive Population III stars are hot enough to significantly ionize it. He II Balmer-α photons (with a rest wavelength of 1640 Å) are therefore the most promising diagnostic of such massive stars: they are produced through recombination cascades following the ionization of He II. See §5.4 for a discussion of this signature.
3. *Metal lines:* In nearby galaxies, many metal lines offer diagnostics of ISM characteristics like the density, metallicity, and temperature of the nebulae surrounding star-forming regions. As instruments improve, these will no doubt be just as useful for measurements of high-z galaxies, although (with most of the lines having rest wavelengths in the optical) they are less accessible for the more distant sources.

10.3.2 Gravitational Lensing and the Luminosity Function

At low and moderate redshifts, the measured luminosity function can generally be taken as the intrinsic luminosity function, with the interesting physical interpretation dependent on understanding the mapping from luminosity (in the waveband of interest) to the constituents of the galaxy, its host dark matter halo, and its large-scale environment. However, at $z > 6$, an additional complication arises: gravitational lensing. The measured luminosity functions are so steep, and the source distance is so large that lensing may substantially affect the observed luminosity function. If so, its interpretation will require an important extra step. We describe a simple model of this lensing effect here, making use of the derivations in §8.10.

insert comma / delete comma / PE

A SIS lens has the simple property that the deflection angle $\hat{\alpha}$ is independent of the impact parameter of the light ray. The condition for multiple imaging (and hence strong lensing) is then that the source lie inside the Einstein radius. The probability that a line of sight to a source at a redshift z_s passes within the cross-sectional area associated with the Einstein radius of SIS lenses, $\pi\theta_E^2$, gives a lensing optical depth

$$\tau_{\rm lens}(z_{\rm s}) = \frac{16\pi^3}{H_0}\int_0^{z_s} dz \frac{D^2(1+z)^2}{(\Omega_m(1+z)^3+\Omega_\Lambda)^{1/2}}\int_0^\infty d\sigma_v \frac{dn}{d\sigma_v}\sigma_v^4, \qquad (10.9)$$

where $(dn/d\sigma_v)\,d\sigma_v$ is the (redshift-dependent) comoving density of SIS halos with a one-dimensional velocity dispersion between σ_v and $\sigma_v + d\sigma_v$. This result is analogous to the halo mass function, but defined with reference to the velocity dispersion rather than the halo mass. The two can be related given a model for the structure of galaxies, as described in chapter 3.

In calculating the probability of lensing it is important to allow for various selection effects. Lenses magnify the observed flux, and lift sources that are intrinsically too faint to be observed over the detection threshold. At the same time, lensing increases the solid angle within which sources are observed, so that their number density in the sky is reduced. If there is a large reservoir of faint sources, the increase in source number due to the apparent brightening outweighs their spatial dilution, and the observed number of sources is increased due to lensing. This *magnification bias* can substantially increase the probability of lensing for source populations whose number-count function is steep, and thus affect the interpretation of the observed luminosity function. The magnification bias for sources at redshift z_s with luminosities between L and $L + dL$ is

$$B(L) = \frac{1}{\phi(L)} \int_{\mu_{\rm min}}^{\mu_{\rm max}} \frac{d\mu}{\mu} \frac{dP}{d\mu} \phi\left(\frac{L}{\mu}\right), \tag{10.10}$$

where $\phi(L)$ is the luminosity function, and $(dP/d\mu)\,d\mu$ is the probability for magnification between μ and $\mu + d\mu$ within the range $\mu_{\rm min} < \mu < \mu_{\rm max}$. For example, the brighter SIS image has a magnification distribution $(dP/d\mu) = 2(\mu - 1)^{-3}$ for $2 < \mu < \infty$.

A simple model for the redshift evolution of SIS lenses uses the mass function of dark matter halos that was derived in §3.4 and identifies $\sigma_v = V_c/\sqrt{2}$ at the virial radius. Another simplified approach is to adopt the observed $(dn/d\sigma_v)$ at $z = 0$ and assume no evolution in the comoving density of lenses. The latter approach gives the approximate results shown in Figure 10.13.

10.4 The Statistics of Galaxy Surveys

Measurements of the statistical properties of galaxies are challenging, and in this section we discuss strategies to constrain their properties, including "one-point functions" like the luminosity or stellar mass functions as well as spatial correlations. The former generally provide insight into the baryonic physics of galaxy formation—how dark matter halos accrete gas and transform it into stars—while clustering provides insight into the dark matter halos themselves.

10.4.1 Estimates of Galaxy Clustering

We have already described our primary tool for understanding the spatial distribution of galaxies, the power spectrum, in §3.6.4, where we developed it through the halo model. In this representation, it contains two terms: the *two-halo* term, which describes the correlations between distinct dark matter halos,

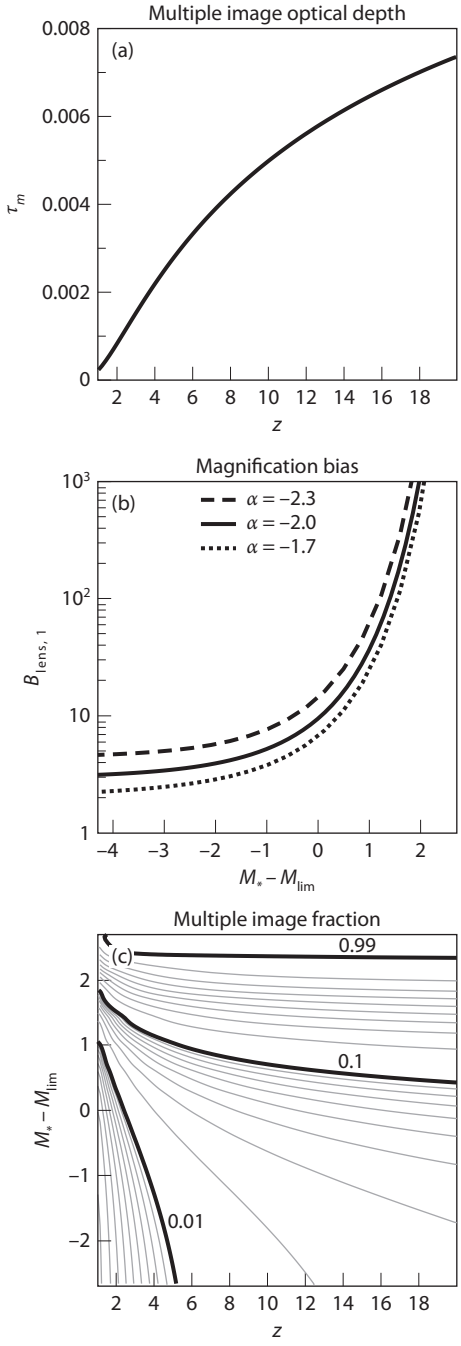


Figure 10.13 Probability for multiple imaging of high-redshift galaxies by an unevolving population of SIS lenses. *(a)* Lensing probability τ_m for obtaining multiple images as a function of source redshift. *(b)* Magnification bias

and the *one-halo* term, which describes the distribution of galaxies within an individual dark matter halo. At high redshifts, where halos are small and probably host only one "galaxy" (at least as we would define them observationally), the latter likely disappears in most halos. Thus, in the language of §3.6.4, we adopt $\langle N|m\rangle = 1$ for all halos above a minimum mass set by accretion or feedback, $m_{\rm min}$, and zero otherwise, and we set the variance in that distribution to zero.

The key additional necessary input is some way of mapping the luminosity (in some photometric band or emission line) or other observable to halo mass. One framework for doing so is the *conditional luminosity function*, in which a function $p(L|m)$ parameterizes the probability distribution of galaxy luminosity as a function of mass. This function can be very complex, of course, but for now we simply assume that it can be constructed from some theoretical or phenomenological arguments. As a crude example we could suppose that galaxies are luminous for a fraction $f_{\rm duty}$ of the time and that there exists a one-to-one relationship between luminosity and mass, $L(m)$, while they are "on." In that case,

$$p(L|m) = (1 - f_{\rm duty})\delta(0) + f_{\rm duty}\delta(L[m]). \tag{10.11}$$

Assuming only one galaxy per halo, and given a minimum observable luminosity $L_{\rm min}$, the predicted galaxy power spectrum will be

$$P^{\rm gal}(k, z) = P_{\rm lin}(k, z)\left[\int dm\, \frac{f(>L_{\rm min}|m)}{\bar{n}_{\rm obs}} n(m, z) b_{\rm eff}(k|m, z)\right]^2, \tag{10.12}$$

where

$$f(>L_{\rm min}|m) = \int_{L_{\rm min}}^{\infty} dL\; p(L|m). \tag{10.13}$$

Comparison with equation (3.84) shows that we have dropped the halo profile (under the assumption that each halo contains only one galaxy) and replaced $\langle N|m\rangle$ with $f(>L_{\rm min}|m)$, which is the probability that a halo of mass m hosts an observable galaxy—clearly, these are two different ways of expressing the occupation fraction of dark matter halos. We can then define a mean bias for this galaxy sample, $\bar{b}_{\rm eff}(k)$, by averaging over all the observable galaxies, so that $P_{\rm gal,obs} \approx \bar{b}_{\rm eff}(k)^2 P_{\rm lin}(k, z)$; in the limit of linear fluctuations, this mean bias is independent of scale and can be predicted using the excursion set formalism via equation (3.54).

Figure 10.13 (Continued.) as a function of the difference between the characteristic magnitude of a galaxy $M_\star$ (assuming a Schechter luminosity function) and the limiting survey magnitude $M_{\rm lim}$. Three values of the faint-end slope of the luminosity function (labeled by α here) are shown. *(c)* Contours of the fraction of multiply imaged sources as a function of source redshift and $(M_\star - M_{\rm lim})$, assuming a faint-end slope for their luminosity function of −2. Wyithe, J. S. B., et al., *Nature* **469**, 7329 (2011). Copyright 2011 by Nature Publishing Group.

Because this effective bias depends on the underlying mass of the galaxy halos—a property of the population that is otherwise nearly impossible to measure—the galaxy power spectrum is of fundamental observational importance. We therefore next describe how it can be measured and the errors that arise when doing so.

Let us suppose that we have a survey over some finite volume V. For now we assume that the three-dimensional locations of the galaxies are known through some spectroscopic survey. Let us define the galaxy overdensity for a mode k_i through the Fourier transform of the galaxy density field,

$$\delta_g(\mathbf{k}_i) = \int \frac{d^3x}{V} W_g(\mathbf{x}) \delta_g(\mathbf{x}) e^{i\mathbf{k}_i \cdot \mathbf{x}}, \tag{10.14}$$

where $W_g(\mathbf{x})$ is the *window function*, which is nonzero only inside the survey volume and is normalized so that $\int d^3x W_g(\mathbf{x}) = V$, the total survey volume.[ii] We then write the observed power spectrum of the galaxy distribution as

$$\left\langle \delta_g(\mathbf{k}_i) \delta_g(\mathbf{k}_j) \right\rangle = \mathbf{C}^S_{ij} + \mathbf{C}^N_{ij}, \tag{10.15}$$

where $\mathbf{C}^S$ is the *signal covariance matrix*, and $\mathbf{C}^N$ is the *noise covariance matrix*. Here the angular brackets denote an average over the density modes in the Universe.

If we substitute equation (10.14) for $\left\langle \delta_g(\mathbf{k}_i) \delta_g(\mathbf{k}_j) \right\rangle$, this average operates on the galaxy overdensity terms, and from equation (2.16), becomes the correlation function of the galaxy field:

$$\mathbf{C}^S_{ij} = \frac{1}{V^2} \int d^3x\, d^3x'\, W_g(\mathbf{x}) W_g(\mathbf{x}') \xi^{\rm gal}(\mathbf{x} - \mathbf{x}') e^{i\mathbf{k}_i \cdot \mathbf{x}} e^{-i\mathbf{k}_j \cdot \mathbf{x}'}. \tag{10.16}$$

However, the correlation function is simply the Fourier transform of the power spectrum (see equation 2.28), so we can write

$$\mathbf{C}^S_{ij} = \int \frac{d^3k}{(2\pi)^3} P^{\rm gal}(k) \frac{\tilde{W}_g(\mathbf{k}_i - \mathbf{k}) \tilde{W}^*_g(\mathbf{k}_j - \mathbf{k})}{V^2}, \tag{10.17}$$

where $\tilde{W}_g(\mathbf{k}')$ is the Fourier transform of the window function. Comparison with equation (2.32) shows that this result is closely related to the variance in the fluctuations within the survey volume.

To gain some intuition for this expression, let us consider some concrete choices for the window function. First, suppose that we observe a spherical volume of radius R around some central point $\mathbf{x}_0$.[iii] Then,

$$\tilde{W}_g(\mathbf{k}_i - \mathbf{k}) = \int_{|\mathbf{x} - \mathbf{x}_0| < R} d^3x\, e^{i(\mathbf{k}_i - \mathbf{k}) \cdot \mathbf{x}}. \tag{10.18}$$

[ii] In practice, we can incorporate any other selection function, such as the likelihood of detecting a galaxy at a particular redshift given some photometric selection criterion, into the window function. Such practical considerations make the analysis more complex but do not change the basic methodology.

[iii] Choosing the central point to be the observer would correspond to a volume-limited sample of galaxies around us. However, for high-redshift surveys, the center of the survey would naturally lie at some distant point.

insert comma (PE) / (2.15) (from renumbering)

2.19 / from renumbering

(2.31) / from renumbering

In the limit that $R \to \infty$, this integral is proportional to a Dirac delta function, so we have

$$\mathbf{C}^S_{ij} \approx (2\pi)^3 P(k_i)\delta^D(\mathbf{k}_i - \mathbf{k}_j), \tag{10.19}$$

2.16 / from renumbering

which matches our original definition of the power spectrum (equation ~~2.17~~). A finite R broadens the delta function, so that the Fourier transform $\tilde{W}_g$ has a nonzero width $\sim(2\pi)/R$. This means that the measured signal will be a weighted average of all modes with $|\mathbf{k} - \mathbf{k}_i| < (2\pi)/R$. Modes with wavelengths larger than the survey volume will be unobservable—they have such small k as to be washed out; those with $k \gg 1/R$ will be essentially unaffected.

At least for existing and near-future observatories, a more realistic survey geometry is a "pencil beam": a narrow angular region (a few arcminutes across for HST or JWST) with a large depth in the radial direction, corresponding perhaps to Lyman-break selection within a particular filter set. In that case, the volume may reasonably be approximated as a long box with radial depth Δz and transverse widths Δx and Δy, such that $\Delta z \gg \Delta x,\ \Delta y$. For a rectangular box, the window function is

$$\tilde{W}(\mathbf{k}) = \tilde{W}_{\Delta x}(k_x)\tilde{W}_{\Delta y}(k_y)\tilde{W}_{\Delta z}(k_z), \tag{10.20}$$

with (k_x, k_y, k_z) the Cartesian components of the wave vector and

$$\tilde{W}_{\Delta x}(k_x) = \frac{\sin(k_x \Delta x/2)}{(k_x \Delta x)/2}, \tag{10.21}$$

and similarly for $\tilde{W}_{\Delta y}$ and $\tilde{W}_{\Delta z}$. This function also is ~ 1 for $k_x \ll \pi/\Delta x$ and falls off at $k_x \sim \pi/\Delta x$. The anisotropy of the window means that the mode sampling depends on their orientations. Modes transverse to the line of sight must have only $k_{x,y} \ll 2\pi/\Delta x$ to be sampled cleanly, but modes along the line of sight must have only $k_z \ll 1/\Delta z$. Even these modes, however, are subject to aliasing from short-wavelength transverse modes, much like the Lyman-α forest power spectrum discussed in §4.3.4, which degrades the ability of such surveys to measure large-scale power.

greek lowercase pi to replace 1 / AA

delete 2 / AA

The noise term $\mathbf{C}^N$ arises from the finite number of galaxies. This *shot-noise* term is inevitable in any experiment that samples a discrete population of objects. Let us assume that the number of galaxies within a given volume obeys Poisson statistics with the mean expected count $\bar{N}$ determined by the underlying density field. The probability of finding N galaxies in a region is then $p(N) = \bar{N}^N e^{-N}/N!$, with $\langle N \rangle = \bar{N}$, and $\langle N^2 \rangle = \bar{N}(\bar{N} + 1)$. For this discrete shot-noise component, the average in equation (10.15) becomes $\langle \delta_i \delta_j \rangle = \bar{N}^{-2}\langle (N_i - \bar{N})(N_j - \bar{N}) \rangle = \bar{N}^{-1}$ if $i = j$ and zero otherwise. This expression replaces the power spectrum in equation (10.17). Finally, we assume that we can choose regions sufficiently small so that each one is either empty or contains at most one galaxy; in that case $\bar{N} = \bar{n}$, the galaxy number density. Finally, by analogy with equation (10.17), we define the *shot-noise power spectrum* as $P^{\rm shot}(k) = 1/\bar{n}$, or its dimensionless form

$$\Delta^2_{\rm shot}(k) = \frac{k^3}{2\pi^2 \bar{n}}. \tag{10.22}$$

Shot noise is an inevitable source of error in any galaxy survey; fortunately, provided one has a good estimate for $\bar{n}$, it can be accurately removed. Shot noise therefore poses a significant problem only when $\bar{n}$ is small, for example, if the survey targets extremely bright, rare galaxies with $L \gg L_\star$.

The power spectrum is by far the most common measure of clustering, owing to the relative ease with which it can be observed. However, it measures only the variance (as a function of scale) of the underlying distribution; higher-order correlations, like skewness, must be measured in other ways. A particularly simple approach to test for these is with *counts in cells*, in which one divides the survey volume into small cells and examines the distribution of galaxy counts in the cells.

Another complication arises if the radial locations of the galaxies are not available, for example, if the galaxies are found through the Lyman-break technique without precise redshifts. In that case, clustering can still be measured along the plane of the sky. This *angular correlation function* (or its counterpart, the angular power spectrum) was for many years the best measure of clustering, even at low redshifts. Intuitively, the angular correlation function is simply the projection of the three-dimensional form onto the plane of the sky. For small angular separations, this is easy to do; we discuss it in more detail in §13.2.

We also note that whenever redshift is used as a proxy for distance (as in a spectroscopic galaxy survey), peculiar velocities in the galaxy field distort the redshift-distance mapping. Fortunately, the velocity effects can be isolated, because they do not affect positions across the plane of the sky: we therefore expect a difference in the clustering measured along the line of sight and along the plane of the sky. Because these peculiar velocities themselves trace the underlying matter distribution, the corresponding *redshift–space distortions* can provide unique information about it. We discuss them further in §12.5.1.

10.4.2 Measuring the Luminosity Function

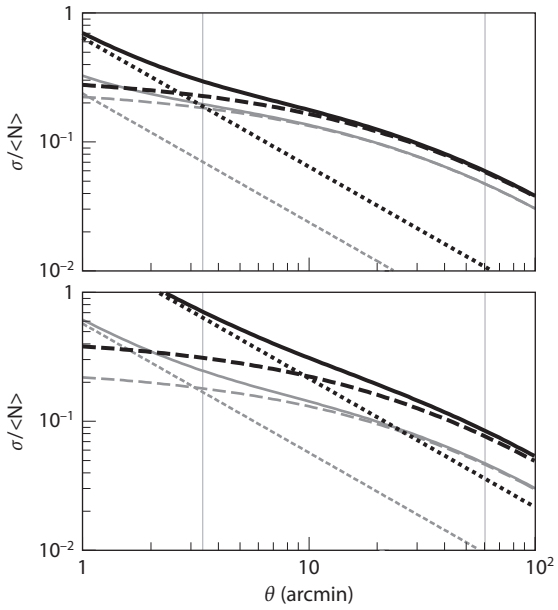
In addition to being of intrinsic interest as a measure of halo mass, clustering also affects the statistical uncertainty in the number counts of galaxies within surveys of limited volume, the *cosmic variance*.[iv] This is crucial to understand for estimates of luminosity and stellar mass functions, because it determines their precision. By analogy with equation (~~2.32~~ 8), the fractional variance in an estimate of the galaxy number counts is the integral of the signal and noise power spectra over all k-modes, weighted by the survey window function:

(2.31 / from renumbering

$$\frac{\sigma^2_{\rm tot}}{\langle N\rangle^2} = \int \frac{dk}{k}\left[\Delta^2_{\rm gal}(k) + \Delta^2_{\rm noise}(k)\right]\frac{|W_g(\mathbf{k})|^2}{V^2}. \qquad (10.23)$$

The cosmic (or sample) variance, which is the first term on the right-hand side of equation (10.23), results because the survey field sometimes lies in a region of high galaxy density and sometimes lies in an underdense region or a void.

[iv] Formally, "cosmic variance" refers to the residual errors from the finite volume of our entire Universe, not simply the observed field (which determines the *sample variance*). In practice, however, the two terms are used almost interchangeably.



dark / AA

light / AA

Figure 10.14 The theoretically predicted contributions to the total variance (equation 10.23; solid lines) in Lyman-break galaxy dropout surveys as a sum of cosmic variance (dashed lines) and Poisson shot noise (dotted lines) contributions. The top and bottom panels show results for surveys extending over the ranges $z = 6$–8 and $z = 8$–10, respectively. The thin lines assume a luminosity threshold of $z_{850,\mathrm{AB}} = 29$, while the thick ones make a cut at $z_{850,\mathrm{AB}}$=27, where here $z_{850,\mathrm{AB}}$ refers to the AB magnitude in the z_{850} photometric band. Figure credit: Munoz, J., Trac, H., & Loeb, A. *Mon. Not. R. Astron. Soc.* **405**, 2001 (2010). Copyright 2010 by the Royal Astronomical Society.

Figure 10.14 compares the contributions from cosmic variance and shot noise as calculated by linear theory for a mock survey as a function of its opening angle, $\theta = r_t/d_A(z)$, where r_t is the transverse width of the survey. This plot can be used to estimate the effectiveness of future surveys with large fields of view. Here we have used a simple model to assign luminosities to dark matter halos, taking $f_{\mathrm{duty}} = 0.25$ and a star formation efficiency $f_\star = 0.16$. Note that the shot noise is important only on small scales, even though the fluctuations from gravitational clustering also decrease with the opening angle of the survey.

According to linear theory, the probability distribution of the count of galaxies is Gaussian with variance given by the sum of the cosmic and Poisson components, so the power spectrum provides a complete representation. However, nonlinear evolution in the matter field induces non-Gaussian structure; because bright high-redshift galaxies are so rare, these nonlinearities can have important effects. Figure 10.15 shows these deviations in the context of a

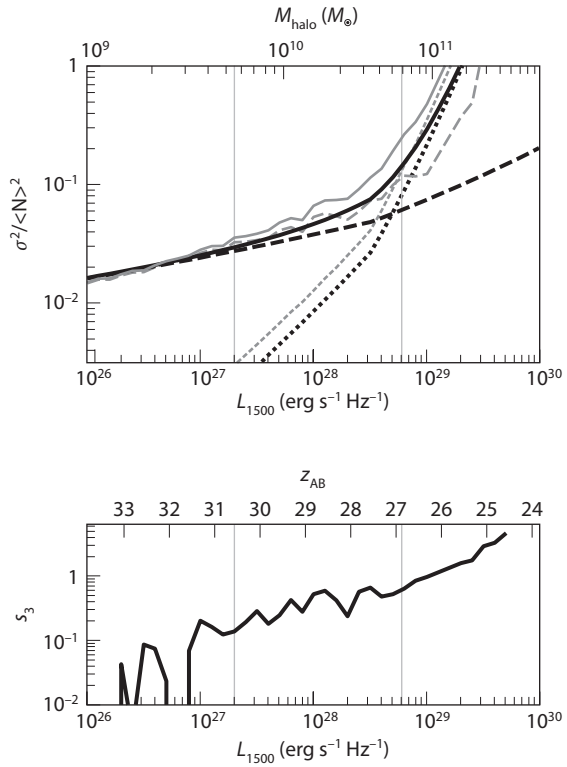


Figure 10.15 *Upper panel:* Predicted relative contributions to the fractional variance in the number counts of galaxies as a function of UV luminosity at an emission wavelength of 1500 Å within a Lyman-break dropout survey spanning the redshift interval $z = 6$–8 with a $3.4' \times 3.4'$ field of view (matching the HUDF and approximately that of JWST). Solid lines show the total variance, while long-dashed and dotted lines show the contributions from cosmic variance and Poisson noise, respectively. The upper curves show the results from numerical simulations, while the lower curves were calculated analytically based on linear perturbation theory. Vertical lines bracket the region where the variance is higher than expected due to the skewness of the full galaxy count probability distribution but is not Poisson dominated. The middle and top horizontal axes translate the monochromatic luminosity to z-band AB magnitude and host halo mass, respectively. *Lower panel:* Skewness of the full galaxy count probability distribution calculated from a numerical simulation based on equation (10.24). Munoz, J., Trac, H., & Loeb, A. *Mon. Not. R. Astron. Soc.*, **405**, 2001 (2010). Copyright 2010 by the Royal Astronomical Society.

light / AA

dark / AA

pencil-beam survey of galaxies (with a $3.4' \times 3.4'$ field of view, as for the HUDF) in the redshift range $z = 6$–8. When compared with numerical simulations, the galaxy count statistics are well approximated by the linear-theory expressions at low luminosity limits.

However, for brighter galaxies, linear theory begins to fail. The upper solid curve in the top panel of Figure 10.15 shows the variance calculated from numerical simulations that include nonlinear evolution. These are larger than the analytic prediction (shown by the lower solid curve) for halo masses $M_h > 10^{10}\,M_\odot$. With the simulations, one can investigate how this happens. The *skewness* is

$$s_3 = \frac{\left\langle (N - \langle N \rangle)^3 \right\rangle}{\sigma^3}. \tag{10.24}$$

The skewness as a function of minimum luminosity is presented in the bottom panel of Figure 10.15. It is large at $M_h > 10^{10}\,M_\odot$ (the seemingly large amplitude variations in the skewness at low luminosity for $z = 6$–8 are due to small numerical fluctuations around the near-zero skewness from numerical simulations, plotted on a log scale). The numerical simulations indicate that the probability distribution of massive halos (and hence presumably bright galaxies) has a non-Gaussian shape. Deviations between the analytic and numerical 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.

10.4.3 Measuring the Galaxy Power Spectrum

We have now shown how to estimate the galaxy power spectrum and how its fluctuations affect number counts of galaxies (and hence the luminosity function). As a final step, let us briefly discuss the errors on a measurement of the power spectrum itself: *How large a survey is necessary to reliably measure the clustering of a galaxy sample?*

Given that real galaxy surveys have complex selection functions, the best way to answer this question is ultimately with detailed simulations of the survey strategy. The next best way is with the *Fisher information matrix*, which provides a robust lower limit to the errors on a given set of parameters in any experiment. We consider this latter approach here. Suppose one wishes to measure the amplitude of the power spectrum over a range of wavenumbers $(k, k + \Delta k)$ in a survey of volume V (these are known as *band powers*). Ignoring boundary effects from the finite survey volume, the minimal error on the band powers is[11]

$$\frac{\Delta P(k)}{P(k)} = (2\pi)\sqrt{\frac{2}{k^2 \Delta k \Delta\mu V}}\left[\frac{1 + \bar{n}P(k)}{\bar{n}P(k)}\right]. \tag{10.25}$$

This expression is straightforward to understand. Recall that the power spectrum quantifies the variance in the density field among a set of modes. Suppose we have N independent estimates of these mode amplitudes. From elementary statistics, the mean squared error on an estimate of the variance from this dataset will be $(\Delta P)^2 \sim \sigma_P^2/N$, where σ_P^2 is the variance of the measurements: in our case, $\sigma_P = P + \bar{n}^{-1}$. We thus need determine only the number of independent samples of the density field in a given power spectrum bin. First,

let us write the Fourier space volume of a binned measurement of the power spectrum as $2\pi k^2 \Delta k \Delta\mu$, where μ is the cosine of the angle between the bin's central wave vector and the line of sight. (As mentioned briefly in §10.4.1 and more extensively in §12.5.1, peculiar velocities induce an anisotropy with respect to μ. For a crude measurement, however, we can average over all modes with a single amplitude, so that $\Delta\mu = 2$, and the volume corresponds to a spherical shell in k-space.)

The final question is how many samples lie within this Fourier space volume. Recall from §10.4.1 that the finite survey volume mixes all modes closer together than $\sim(2\pi)^3/V$. However, the reality of the density field imposes a constraint on its Fourier transform, relating pairs of modes with $\mathbf{k}$ and $-\mathbf{k}$ to each other. Thus, the number of *independent* samples is $N \approx 2\pi k^2 \Delta k \Delta\mu \times [V/(2\pi)^3] \times 1/2$. The prefactor in equation (10.25) is simply $1/\sqrt{N}$.

This approach provides a reasonable estimate for the volume required to measure galaxy bias from a survey. In the regime where shot noise is unimportant (i.e., $P \gg \bar{n}^{-1}$), a measurement with 10% precision requires a volume of $\sim 10^7 (k/0.1\ \mathrm{Mpc}^{-1})^{-3}\ \mathrm{Mpc}^3$. High-$k$ modes evidently do not require particularly large volumes, but surveys run into shot-noise limitations unless they go very deep; even the faintest HUDF galaxies have $\bar{n} \sim 0.01\ \mathrm{Mpc}^{-3}$ (see Figure 10.9). Shot noise compromises modes with $\bar{n}P < 1$. In that regime, it is generally advantageous to construct a deeper, rather than wider, survey.

Chapter Eleven

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

Early in the book, we explored the physics behind structure formation, which led to the first sources of light in the Universe. In chapter 10, we began applying this framework to observable systems: galaxies. We have now arrived at the point where we can study a number of specific observational probes of the high-redshift Universe. We begin that endeavor in this chapter by examining the Lyman-α line, an extraordinarily rich and useful—albeit complex—probe of both galaxies and the IGM. In the next two chapters, we will describe a variety of other observables.

11.1 Lyman-α Emission from Galaxies

We saw in §10.2.1 that young star-forming galaxies can produce very bright Lyman-α emission;[1] indeed searching for such line emitters is one of the most effective ways to find high-z galaxies. Although the radiative transfer of these photons through their host galaxies is typically very complex, a good starting point is a simple model in which a fraction of stellar ionizing photons are absorbed within their source galaxy, forming embedded H II regions. The resulting protons and electrons then recombine, producing Lyman-α photons. Assuming ionization equilibrium, the rate of these recombinations must equal the rate at which ionizing photons are produced. However, about one-third of these recombinations do not cascade through the Lyman-α transition and so do not contribute to this line.[i]

Because only hot, massive stars—which live for only several million years—produce ionizing photons, it is a good approximation to assume that the rate at which any given galaxy generates these photons is proportional to its instantaneous star formation rate $\dot{M}_\star$. The proportionality constant depends on the initial mass function (IMF) of stars, because that determines what fraction of stellar mass enters these massive, hot stars. For example, a galaxy with a constant star formation rate (SFR), a Salpeter IMF, a metallicity $Z = 0.05\ Z_\odot$, and no binary stars produces $\dot{Q}_i = 4.3 \times 10^{53}$ ionizing photons per second per year per solar mass of stars formed.[2] However, a top-heavy Population III IMF has an order-of-magnitude larger ionizing efficiency.

delete overdot / AA

[i] Note also that direct recombinations to the ground state (which also occur approximately one-third of the time, from the ratio of the case-A and case-B recombination coefficients α_A and α_B in equations 4.17 and 4.18) simply regenerate the initial ionizing photon, so they do not contribute to the net balance. We therefore worry only about those ionizing photons produced directly by the stars.

Finally, if we assume, as usual, that a fraction $f_{\rm esc}$ of the ionizing photons escape their host galaxy, then the intrinsic line luminosity of a galaxy is

$$L_{\rm Ly\alpha}^{\rm int} = \frac{2}{3}\dot{Q}_i h\nu_\alpha (1 - f_{\rm esc})\dot{M}_\star. \tag{11.1}$$

For context, a Salpeter IMF from 1 to 100 $M_\odot$ with $Z = 0.05 Z_\odot$ has a prefactor $4.4 \times 10^{42}(1 - f_{\rm esc})$ erg s^{-1}, if the star formation rate $\dot{M}_\star$ is measured in solar masses per year.

Unfortunately, inferring physical properties about distant galaxies from the Lyman-α line is complicated not only by the uncertain factors $f_{\rm esc}$ and $\dot{Q}_i$ but also by the radiative transfer of these line photons through the interstellar and circumgalactic medium of each galaxy as well as the more distant IGM. Because the Lyman-α line is so optically thick in these environment, line photons scatter many times before they can escape the galaxy, and once they leave it, they can be scattered away from the line of sight and vanish. This scattering can change not only the overall brightness of the line but also its frequency structure and relation to the galaxy's continuum photons. The *observed* line luminosity is then

$$L_{\rm Ly\alpha}^{\rm obs} = \frac{2}{3} T_{\rm Ly\alpha}^{\rm IGM} T_{\rm Ly\alpha}^{\rm ISM} \dot{Q}_i h\nu_\alpha (1 - f_{\rm esc})\dot{M}_\star, \tag{11.2}$$

where $T_{\rm Ly\alpha}^{\rm ISM}$ is the fraction of Lyman-α photons that are transmitted through the galaxy's ISM, and $T_{\rm Ly\alpha}^{\rm IGM}$ is the fraction transmitted through the IGM. We will see later that the latter factor can inform us about the properties of the IGM and reionization.

11.1.1 Radiative Transfer of Lyman-α Photons through the Interstellar Medium

We first consider radiative transfer within a galaxy and its immediate environs; we defer discussion of IGM scattering until §11.2. The important difference from continuum transfer is that line photons can scatter many times (changing both their direction and frequency) as they traverse the ISM. In the case of Lyman-α photons, scattering cannot destroy them (unless collisions mix the $2s$ and $2p$ electron states, which requires extremely high densities) but dust absorption can. Depending on the geometry of the ISM, the increased path length can increase or decrease the brightness of the line relative to the continuum.

Some simple toy models of radiative transfer help develop some intuition for this situation. We generally assume that the absorption cross section follows the Voigt profile, $\sigma_\alpha(v) = \sigma_0 \phi_V(v)$, which allows for both thermal effects (which cause Gaussian broadening in the core of the line) and natural broadening in the wings (which arise from the finite lifetime of the upper state). The full Voigt profile may be written as a convolution of these two mechanisms,

$$\phi_V(v) = \int_{-\infty}^{\infty} p_M(v) L[v(1 - v/c)]\, dv, \tag{11.3}$$

insert period / PE

delete overdot / AA

insert closing parenthesis / PE

where the integral is over the line-of-sight thermal velocities of the particles. Here $p_M(v)$ is the Maxwell-Boltzmann distribution at a temperature T,

$$p_M(v)\,dv = \frac{1}{\sqrt{\pi b^2}} e^{-v^2/b^2}\,dv, \tag{11.4}$$

where $b = \sqrt{2k_B T/m_p}$ is the thermal velocity of the atom.[ii] L is the natural line profile, which for Lyman-α is given by equation (4.8). This is often approximated by a Lorentzian function,[iii]

$$L(v) \approx \frac{1}{\pi} \frac{\gamma}{(v - v_\alpha)^2 + \gamma^2}, \tag{11.5}$$

where $\gamma = A_{21}/(4\pi)$ is the decay constant. In this approximation, the Voigt profile can be written as

$$\phi_V(x)\,dx = \frac{1}{\sqrt{\pi}} \left(\frac{v_\alpha}{v}\right) V(x)\,dx, \tag{11.6}$$

where $x = (v - v_\alpha)/v_D$ is the normalized frequency, with a Doppler broadening $v_D/v_\alpha = b/c$, and

$$V(x) = \frac{A(x)}{\pi} \int_{-\infty}^{\infty} dy \frac{e^{-y^2}}{[B(x) - y]^2 + A^2(x)}. \tag{11.7}$$

Finally, $A(x) = (\gamma/v_D) \times (v_\alpha/v)$, and $B(x) = (v/v_\alpha)x$. A low-order approximation to the Voigt function makes the line structure apparent:

$$V(x) \approx e^{-B^2} + \frac{1}{\sqrt{\pi}} \frac{A}{A^2 + B^2}. \tag{11.8}$$

We are particularly interested in the profile far from the line core (where $B \gg 1$). There, $\sigma_\alpha \approx a/(\sqrt{\pi}x^2)$, where $a \equiv \gamma/v_D = 4.72 \times 10^{-4} T_4^{-1/2}$, and $T_4 \equiv (T/10^4\ \mathrm{K})$ Figure 11.1 shows the absorption cross section for absorbing gas with $T = 10^4$ K (including the full line broadening); note the Gaussian core with width $\sim 10\,\mathrm{km\ s^{-1}}$ and the much weaker, but broader, damping wings.

With these fundamental parameters established, let us consider how line photons can escape from several toy models of gas clouds. We describe the physics quantitatively here and illustrate it in Figures 11.2 and 11.3.

- *Homogeneous, static H I slab, moderately optically thick* (Figure 11.2a): First, consider a Lyman-α photon produced inside a homogeneous static medium of pure H I, with total line-center optical depth $\tau_0 \gg 1$. Because τ is proportional to distance in the medium, we can (and usually will) use it as a proxy for physical location within the system. So long as the photon remains in the Doppler core of the line, it barely diffuses spatially before being scattered by an atom. When a line photon of frequency $x_{\rm in}$ is absorbed

[ii] The Doppler parameter can include turbulence as well by the addition of the turbulent velocity in quadrature.

[iii] However, in many cosmological applications the optical depth can be so high that the asymmetry of the full profile is visible.

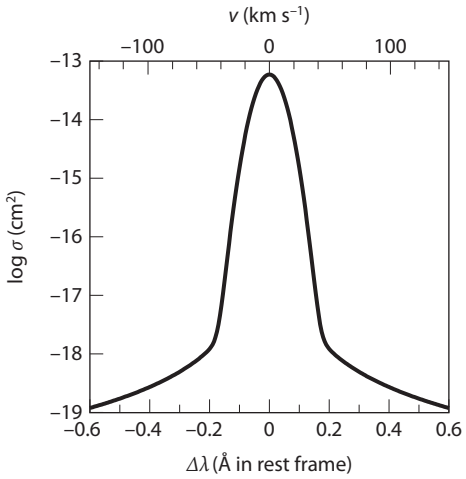


Figure 11.1 Cross section for Lyman-α absorption, as a function of wavelength offset from line center (bottom axis) or velocity difference (top axis). The calculation includes thermal and natural broadening generated by gas with $T = 10^4$ K. Santos, M. R., *Mon. Not. R. Astron. Soc.* **349**, 1137 (2004). Copyright 2004 by the Royal Astronomical Society.

by an atom, it reemits a line photon of the same frequency in its own rest frame. However, in an observer's frame, there is ~~be~~ a net frequency shift determined by the Lorentz transformation between the frames. To linear order, this shift is

PE

$$x_{\rm out} \approx x_{\rm in} - \frac{\mathbf{v}_a \cdot \mathbf{k}_{\rm in}}{b} + \frac{\mathbf{v}_a \cdot \mathbf{k}_{\rm out}}{b} + g(\mathbf{k}_{\rm in} \cdot \mathbf{k}_{\rm out} - 1), \qquad (11.9)$$

where $\mathbf{v}_a$ is the velocity vector of the atom, and $\mathbf{k}_{\rm in}$ and $\mathbf{k}_{\rm out}$ are the propagation directions of the incoming and outgoing photons, respectively. The g in the last term accounts for recoil; it is unimportant here, but we will revisit it in §12.2.2. Typically, the scattering atom will have the same velocity along the photon's direction of motion as the atom that emitted it, but it can have a much larger total velocity. In that case, the scattered photon will be far from the line center, in the damping wings.

If the medium is not too optically thick, so that these damping wings are themselves optically thin, the resulting photon can escape so long as it is produced at a frequency where $\tau(x) < 1$; for $\tau_0 = 10^3$, this corresponds to a value for $x \sim 2.6$. We therefore generically expect that the resulting emission will have a double-peaked profile: photons near linecenter are trapped in the cloud; only when they diffuse to large positive or negative velocity are they able to escape. The Lyman-α surface brightness distribution will also be compact, because photons escape after a small number of scattering events.

insert space / PE

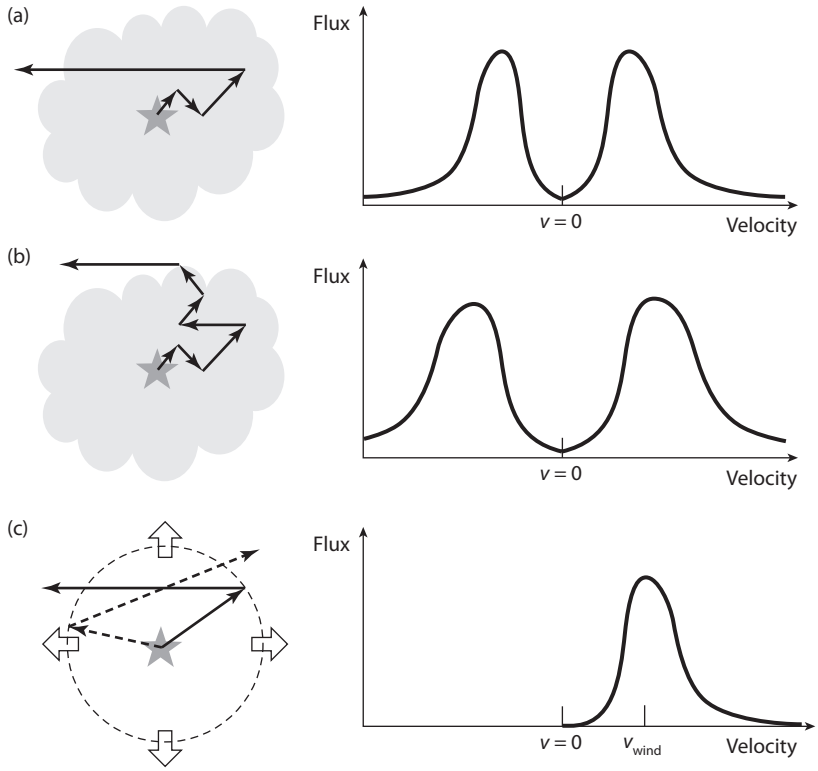


Figure 11.2 Schematic illustrations of the radiative transfer of Lyman-α photons through a dust-free, homogeneous interstellar medium. *(a)* For a moderately optically thick cloud, a Lyman-α photon undergoes a random walk until its frequency diffuses into the wings of the line, escaping once the optical depth $\tau_\nu < 1$ at the photon frequency ν. This produces a double-peaked emission line and spatially compact emission. *(b)* For a very optically thick cloud, the photon may need to diffuse out of the cloud in space rather than frequency. In that case, the two peaks are substantially broader, and the emission is spatially extended. *(c)* If the scattering medium has a velocity gradient (here we imagine an expanding shell of gas), photons initially directed toward the observer are scattered away by the blueshifted near side of the shell. However, photons directed toward the far side of the shell that backscatter toward the observer pass through the near side of the shell far out of resonance and escape toward the observer. The result is a single emission line centered on the velocity of the far side of the shell.

- *Homogeneous, static H I slab, very optically thick* (Figure 11.2(b)): Thus, in a moderately optically thick medium, these escaping photons simply result from rare scatterings off high-velocity atoms. If, however, the damping wings are themselves optically thick, $\tau > 10^3$, the problem is more complicated, though the net result is easy to understand: the photons must make

delete opening parenthesis / PE

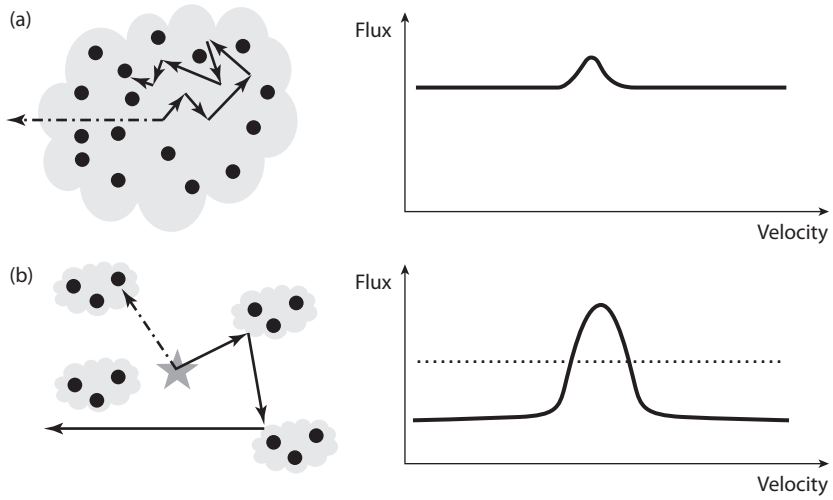


Figure 11.3 Schematic illustrations of the radiative transfer of photons through an interstellar medium with dust. *(a)* If the scattering medium is homogeneous, with dust (black circles) uniformly distributed through it, Lyman-α photons (solid black arrows) undergo a random walk in space and frequency before they can escape (see Figure 11.2). Their long path length implies that they have an increased probability of extinction by dust compared with that of a continuum photon (dot-dashed arrow). The result is suppression of the emission line (shown for simplicity as a single line). *(b)* If the dust is instead confined to discrete, optically thick clouds, the line can instead be enhanced. In this case, Lyman-α photons scatter off the outskirts of each cloud until they escape the system, so they pass through relatively little dust. Continuum photons, in contrast pass through the clouds and so suffer *more* dust extinction.

insert comma / PE

it even farther into the wings to escape. In this regime, a photon in the wings is more likely to scatter off an atom with a small thermal velocity (in its damping profile) than an atom traveling at a matched velocity,

To escape, the photon must then undergo a random walk of repeated scatterings, which occasionally takes it far enough from line center to escape. Because scatterings usually occur in the core, each one induces an rms frequency shift $x \sim 1$, with a small bias $-1/x$ toward returning to line center; a photon thus typically undergoes $N_s \sim x^2$ scattering events before returning to line center. Between scatterings, the photon traverses a path length (in optical depth units) of $\tau\phi_V(x) \sim 1$. Thus, over its entire random walk, it diffuses a distance $\tau_0^{\rm rms} \sim \sqrt{N_s}\tau \sim |x|/\phi_V$. If this distance exceeds the physical size of the system (τ_0 in these units), the photon can escape. In the wings of the line, where $\phi_V \sim a/x^2$, this requires that the photon have a critical normalized escape frequency

$$|x_{\rm esc}| \sim [a\tau_0]^{1/3} \approx 30 T_4^{-1/3} N_{21}^{1/3}, \tag{11.10}$$

where N_{21} is the column density of the system in units of 10^{21} cm^{-2}. Thus, in this highly optically thick case, the photon must scatter far enough in the wings of the line to physically escape the system before scattering back to line center. This condition, combined with the power law form of ϕ_V in the wings, also makes the blue and red emission peaks wider than than in the moderate optical depth case. The surface brightness of the line will be extended even if the source is compact, because photons diffuse spatially as well as in frequency before escaping.

insert comma / PE

- *Homogeneous H I slab, with velocity gradient* (Figure 11.2(c): We next consider a medium with a velocity gradient. Such a gradient can correspond either to expansion, arising from winds (which we believe to be ubiquitous in the star-forming galaxies likely to host Lyman-α emission lines), or contraction, from the infall of surrounding material around the galaxy.

delete opening parenthesis / PE

  First, consider an expanding medium. Then, according to equation (11.9), scattered photons typically acquire a redshift: $\mathbf{v}_a \cdot \mathbf{k}_{\rm in}$ is positive for photons propagating outward, while (assuming isotropic emission) $\langle \mathbf{v}_a \cdot \mathbf{k}_{\rm out} \rangle = 0$, so $x_{\rm out} < x_{\rm in}$ on average. Photons with $x < 0$ are therefore moved farther into the line wings, facilitating their escape, while photons with $x > 0$ are moved back toward line center. So long as the expansion velocity is much larger than the thermal velocities, this prevents photons that experience large positive frequency jumps from escaping. Thus, we expect only a single emission line on the red side. In contrast, in a contracting medium, photons typically acquire a blueshift, producing a single emission line on the blue side.

  In this case the frequency shift of the surviving line depends on the velocity and density structure of the medium. The case of most practical relevance is a wind, in which a large column of H I occurs at $\pm v_{\rm wind}$ along the line of sight, with negligible absorption elsewhere. In this case photons that begin their escape toward the observer (i.e., through the blueshifted wind) are absorbed. After their first scattering, photons that begin their escape toward the far component of the wind lie to the red side of the line. Those that scatter back toward the observer are then far to the red of the (blueshifted) line center of the near component of the wind and can continue to the observer. The observed velocity offset is then $v_{\rm wind}$ and provides a good diagnostic of the wind velocity.

- *Homogeneous H I slab with dust* (Figure 11.3a): Now we can add dust to a (static) medium and see how it can destroy the Lyman-α photons. We let the *total* dust interaction cross section, per hydrogen nucleus, be σ_d; this includes both absorption, with a cross section $\sigma_a = \epsilon_a \sigma_d = (\sigma_{a,21}/10^{-21}\ {\rm cm}^2)$ per hydrogen atom, and scattering (with cross section $\sigma_d - \sigma_a$). For the well-studied dust in the Milky Way, $\sigma_{a,21} \approx 1$, and $\epsilon_a \approx 0.5$; of course, these values will depend on the metallicity and dust formation mechanisms in high-redshift galaxies (see §8.9.2). The average absorption probability per interaction (with either dust or H I) is therefore

$$\epsilon_{\rm dust} = \frac{\sigma_a}{x_{\rm H\ I}\phi_V(x)\sigma_0 + \sigma_d} \approx \frac{\beta}{\phi_V(x)}, \qquad (11.11)$$

where $\beta = \sigma_a/(x_{\rm H\ I}\sigma_0) = 1.69\times10^{-8}T_4^{1/2}\sigma_{a,21}/x_{\rm H\ I}$, and we have assumed that dust interactions are rare compared with H I scattering.

Now, recall that to escape the H I cloud, the photon must first scatter far into the wings of the line and then remain in the wings as it spatially diffuses out of the system. During that process, the photon will scatter N_s times; the probability that it will be absorbed by dust and destroyed is therefore $P_{\rm abs} \sim N_s\epsilon_{\rm dust} \sim x^4\beta/a$ in the damping wing. This probability is near unity if $|x| > x_{\rm abs}$, where

$$x_{\rm abs} \sim (a/\beta)^{1/4} \sim 12.9\left(\frac{x_{\rm H\ I}}{T_4\sigma_{a,21}}\right)^{1/4}. \tag{11.12}$$

A typical photon will therefore be unable to escape if $x_{\rm esc} > x_{\rm abs}$; if the line-center optical depth exceeds a critical value $\tau_c \sim (a\beta^3)^{1/4}$, the emission line will be strongly suppressed. This critical value corresponds to a column density of only $N_{21,c} = 0.08T_4^{1/4}(x_{\rm H\ I}/\sigma_{a,21})^{3/4}$, well below the typical column densities of galaxies (which are comparable to damped Lyman-α absorbers [DLAs]). Thus Lyman-α destruction can be very important inside a dusty ISM. As a rule of thumb, in a uniform medium, line photons are more affected by dust than continuum photons, because the many scatterings suffered by the former force them to follow a much longer path length than continuum photons, thus providing a much larger opportunity for destruction by dust.

- *Multiphase medium with dust* (Figure 11.3b): Finally, we consider a medium in which both the H I and dust are confined to optically thick, discrete clouds separated by a highly ionized, dust-free "intercloud medium." Here, details of the results will clearly depend on the geometry of the system, but some general considerations do apply. First, note that an inhomogeneous medium will allow *more* transmission than a homogeneous slab with identical column density of neutral gas, because of the same arguments we discussed for Lyman-α transmission in an inhomogeneous IGM (see §4.3.2). Moreover, in some cases the line photons can be *less* affected by dust than continuum photons, because the line photons scatter off the *surface* of the clouds, while the continuum photons plow through them and can encounter *more* dust.

  Detailed calculations show that the frequency shift necessary for dust absorption to dominate over resonant scattering in the line wings, $x_{\rm abs}$, is similar in magnitude to the homogeneous case.[3] However, dust was important in the uniform example because Lyman-α photons *needed* to diffuse in frequency to escape the medium.

  This is not the case for a multiphase medium.[4] In this case, Lyman-α photons enter each cloud on their surface and suffer relatively few scattering events inside each cloud before spatially diffusing out.The photons can then travel a large distance before hitting another cloud, and spatial diffusion through the intercloud medium provides most of the impetus toward escape. Thus, dust absorption will be relatively weak provided the typical frequency shift before escape is less than $x_{\rm abs}$.

In this case, photons acquire frequency shifts both from the thermal motions of the scattering atoms and from the velocity dispersion between the absorbing clouds; if the latter is large (as would be the case if most of the dust were buried in dense molecular clouds), it dominates the frequency diffusion, because—just as for a wind— each cloud is so optically thick that in the observer's frame the photon leaves each cloud with a velocity offset corresponding to that cloud's velocity. If the clouds have a large velocity dispersion, then dust absorption within each cloud will dominate over resonant scattering, because the photons will enter each cloud in the wings of the line.

Although each of these toy models is obviously much simpler than a real galaxy, together they illustrate the complexity of the radiative transfer problem and the many parameters that can dramatically affect the Lyman-α line's amplitude and shape, as well as the surface brightness of a line emitter. In general, even discounting uncertainties from IGM transmission, discussed next, the Lyman-α line is typically very difficult to interpret and is not regarded as, for example, a very reliable measure of the star formation rate. However, its extreme brightness in many galaxies makes it such a useful signpost that it is still the subject of intense study.

11.2 The Gunn-Peterson Trough

We now briefly discuss the fate of continuum photons that begin their lives blueward of Lyman-α during the reionization era. These photons will redshift through the IGM; if they should pass through the Lyman-α resonance, they will experience substantial absorption from that gas. The scattering cross section of the H I Lyman-α resonance line is given by equation (4.8), and we have already computed the total optical depth for a photon that redshifts through the Lyman-α resonance as it travels through the IGM, the so-called Gunn-Peterson optical depth in equation (4.11). The most important aspect of this calculation is the enormous overall optical depth in a fully neutral IGM, $\tau_\alpha \sim 6.5 \times 10^5 x_{\rm H\ I}$ at $z \sim 9$. Thus, we expect that before reionization, photons that redshift across the Lyman-α transition will be completely extinguished (and, indeed, the same will be true so far as $x_{\rm H\ I} > 10^{-3}$).

However, not all photons will redshift through the resonance during the reionization era. Suppose that a photon is emitted by a source at a redshift z_s beyond the "redshift of reionization" $z_{\rm reion}$, which for the purposes of this calculation is simply the last redshift along the particular line of sight of interest where $x_{\rm H\ I} = 1$. (Note that this differs from the conventional definition of the end of reionization as the moment of "overlap" between the ionized bubbles; the variations along different lines of sight can themselves contain interesting astrophysical information.) For simplicity, we further assume that $x_{\rm H\ I} = 1$ for all $z > z_{\rm reion}$. The corresponding scattering optical depth of a uniform, neutral

IGM is a function of the observed wavelength $\lambda_{\rm obs}$,

$$\tau_\alpha(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz\, \frac{c\,dt}{dz} n_{\rm H,0}(1+z)^3 \sigma_\alpha\left[\nu_{\rm obs}(1+z)\right]. \qquad (11.13)$$

At wavelengths corresponding to the Lyman-α resonance between the source redshift and the reionization redshift, $(1+z_{\rm reion})\lambda_\alpha \le \lambda_{\rm obs} \le (1+z_s)\lambda_\alpha$, the optical depth is given approximately by equation (4.11). Since $\tau_\alpha \sim 10^5$, the flux from the source is entirely suppressed in this regime. However, photons blueward of this minimum wavelength do not redshift into resonance until *after* reionization is over, so we should see a recovery in the Lyman-α forest at the blue end of the spectra that in principle informs us about reionization. Similarly, the Lyman-β resonance produces another trough at wavelengths $(1+z_{\rm reion})\lambda_\beta \le \lambda_{\rm obs} \le (1+z_s)\lambda_\beta$, where $\lambda_\beta = (27/32)\lambda_\alpha = 1026\,$Å, and the same applies to the higher Lyman-series lines. If $(1+z_s) \ge 1.18(1+z_{\rm reion})$ then the Lyman-α and the Lyman-β resonances overlap, and no flux is transmitted between the two troughs. The same holds for the higher Lyman-series resonances down to the Lyman-limit wavelength $\lambda_{\rm H\,I} = 912\,$Å.

replace with - and close up / PE

At wavelengths shorter than $\lambda_{\rm H\,I}$, the photons may be absorbed when they photoionize atoms of hydrogen or helium, even if they do not redshift into the Lyman-series lines. The bound-free absorption cross section of hydrogen is given by equation (4.14); the appropriate parameters for He II are given in §4.5 as well. A reasonable approximation to the total cross section for a mixture of hydrogen and helium with cosmic abundances in the range of $4h\nu_{\rm H\,I} = 54.4\,{\rm eV} < h\nu < 10^3\,{\rm eV}$ is $\sigma_{bf} \approx \sigma_0(\nu/\nu_{1.8})^{-3}$, where $\sigma_0 \approx 6\times10^{-17}\,{\rm cm}^2$. The frequency factor in the cross section then cancels exactly the redshift evolution of the gas density, and the resulting optical depth depends only on the elapsed cosmic time, $t(z_{\rm reion}) - t(z_s)$. At high redshifts this optical depth is

subscript Roman H I / AA

$$\tau_{bf}(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz\, \frac{c\,dt}{dz} n_0(1+z)^3 \sigma_{\rm bf}\left[\nu_{\rm obs}(1+z)\right]$$

italics / AA

$$\approx 1.5\times10^2 \left(\frac{\lambda_{\rm obs}}{100\,\text{Å}}\right)^3 \left[\frac{1}{(1+z_{\rm reion})^{3/2}} - \frac{1}{(1+z_s)^{3/2}}\right]. \qquad (11.14)$$

The bound-free optical depth approaches unity only in the extreme UV to soft X-rays, around $h\nu \sim 0.1\,$keV, a regime that is unfortunately difficult to observe owing to absorption by the Milky Way galaxy.

Together, these effects imply very strong absorption of nearly all photons that begin blueward of $\lambda_\alpha(1+z_r)$, except for a recovery at very short wavelengths and at the gaps between the Lyman-series troughs (though these will be blanketed by the Lyman-α and other forests just below $z_{\rm reion}$, so even they will be extremely optically thick).

DES: 2 lines under head ok?

11.3 IGM Scattering in the Blue Wing of the Lyman-α Line

We now return to the fate of photons emitted within (or near) the Lyman-α line of a galaxy or quasar. In this case, the relative velocity and broadening of the

line from bulk, thermal, or turbulent motions is very significant, because it determines whether the photons pass through the Lyman-α resonance—and so experience the full Gunn-Peterson absorption—or remain redward of line center, experiencing much less absorption. We also must consider the environment of the source: whether it is embedded in completely neutral gas or in an ionized bubble, and the surrounding velocity field. In this section we focus on photons emitted blueward of, but still near, line center.

11.3.1 Resonant Scattering Inside Ionized Bubbles

lc / PE

Photons that begin slightly blueward of line center redshift into the Lyman-α resonance near their source. In most models, this nearby region will already have been ionized, either by the source itself or by its neighbors (if it is part of a much larger ionized bubble). Thus it may seem that these photons will survive their journey through the IGM.

However, if we recall that $\tau_\alpha > 10^5 x_{\rm H\,I}$ at these redshifts, it is immediately apparent that even in highly ionized media the absorption can be substantial. In practice, the short mean free paths at high redshifts will most likely prevent the gas from becoming extremely ionized. We can estimate the residual ionized fraction inside an H II region in which the comoving mean free path is λ (which can be restricted either by the ionized bubble walls or by LLSs; see §9.5) by assuming ionization equilibrium and a uniform emissivity (or, in other words, that each bubble contains many sources). The equilibrium condition is then $\Gamma n_{\rm H\,I} = \alpha_B n_e n_{\rm H}$, where $\Gamma \approx \epsilon_{\rm ion} \bar{\sigma}_{\rm H\,I} \lambda/(1+z)$, $\epsilon_{\rm ion}$ is the proper emissivity of ionizing photons (by number), and $\bar{\sigma}_{\rm H\,I} \sim 2 \times 10^{-18}$ cm^2 is the frequency-averaged cross section. If we use the simplest model for the ionizing sources, in which the rate of ionizing photon production is proportional to the rate at which gas accretes onto galaxies, we can write (see equation 9.2) $\epsilon_{\rm ion} = \zeta \dot{f}_{\rm coll} n_{\rm H}$. But we also know that $Q_{\rm H\,II} = \zeta f_{\rm coll}/(1+\bar{n}_{\rm rec})$, where $\bar{n}_{\rm rec}$ is the mean number of recombinations per atom. So we can rewrite the ionizing efficiency ζ in terms of the filling fraction of ionized bubbles and solve for the resonant optical depth due to residual neutral gas $x_{\rm H\,I}$ inside the bubble:

$$\tau_\alpha^{\rm res}(\delta) \approx 40 \frac{(1+\delta)^2}{Q_{\rm H\,II}(1+\bar{n}_{\rm rec})} \left(\frac{10\ {\rm Mpc}}{\lambda}\right) \left(\frac{f_{\rm coll}}{df_{\rm coll}/dz}\right), \qquad (11.15)$$

where we have assumed that the IGM is isothermal to compute the recombination coefficient. The factor involving the collapsed fraction is typically of the order of

3 (insert before period) / PE

Clearly, the optical depth for these photons is large in realistic models; note, however, that it is small enough that many of the radiative transfer effects important for photon escape from galaxies (discussed in §11.1.1) can be ignored, so the absorption from each gas parcel will not have a large frequency width.

11.3.2 The Proximity Effect and Quasar "Near-Zones"

DES: 2 lines under head ok?

Equation (11.15) shows that an average location inside an ionized bubble is not likely to be ionized strongly enough to allow significant transmission

before reionization. However, the region immediately outside an ionizing source will be more ionized than average owing to photons from that very source. At moderate redshifts, this "proximity effect" is a useful measure of the ionizing background, and it is a very attractive probe of the reionization era as well.

The profile of the ionization rate around a quasar at moderate redshifts is simple to understand. Suppose there is a uniform metagalactic background with amplitude $\Gamma_{\rm bg}$. The central quasar with a luminosity L_ν produces a specific intensity $J_\nu \propto L_\nu/R^2$, where R is the proper distance from the quasar. Thus, we expect an ionization rate $\Gamma_q \propto 1/R^2$. Assuming ionization equilibrium, we then have

$$\tau(R) \propto [\Gamma_{\rm bg} + \Gamma_{q,0}(R_0/R)^2]^{-1}, \tag{11.16}$$

Where R_0 is a scale length and $\Gamma_{q,0}$ depends on the quasar luminosity. In principle, a simple fit to the absorption profile as a function of distance from the quasar is sufficient for deriving $\Gamma_{\rm bg}$, especially if $\Gamma_{q,0}$ can be estimated from the observed luminosity of the quasar redward of the Lyman-α line. In practice, these estimates are complicated by variations in the Lyman-α forest lines themselves and by the biased environments of quasars: the quasar will induce substantial changes in the radiation field only within a compact "proximity zone" around the quasar where $\Gamma_q > \Gamma_{\rm bg}$. This zone corresponds to

$$R_{\rm prox} = \frac{1.2\Gamma_{12}^{-1/2}}{\alpha+3}\left(\frac{\nu L_\nu}{10^{44}\ {\rm erg\ s^{-1}}}\right)^{1/2} \text{ proper Mpc}, \tag{11.17}$$

where α is the quasar spectral index, and L_ν is evaluated at the H I ionization edge. This places the proximity zone within the overdense environment of the quasar's halo; the increased absorption from this excess gas partially cancels the effect of the increased ionizing background (the $[1+\delta]^2$ factor in equation 11.15), making the proximity effect more difficult to see.

Because the ionizing background is much smaller during the reionization era, it may at first appear that the proximity effect will be easier to observe. However, in reality, the effect is much more difficult to interpret because the IGM is so optically thick. In this situation, the observable pattern near a luminous source will be gradually increasing absorption until saturation is reached. Figure 11.4 shows some examples; the curves there have each been averaged over several independent lines of sight to reduce the scatter from the inhomogeneous IGM. The horizontal dotted line marks 10% transmission.

The key point is that during the reionization era, there are two possible reasons for such saturation to occur. The first is that the source (usually a quasar) is still in the process of ionizing its neutral surroundings. Then, there will be a sharp transition between the highly ionized H II region and the nearly neutral gas at its edge, which will manifest itself as a dramatic increase in the local optical depth. The second is more similar to the classical proximity effect, except that the absorption may saturate long before the local ionization rate reaches the background value, if $\Gamma_{\rm bg} \approx 0$. Because the observed edge of the transmission does not necessarily correspond to the classical proximity zone, this feature

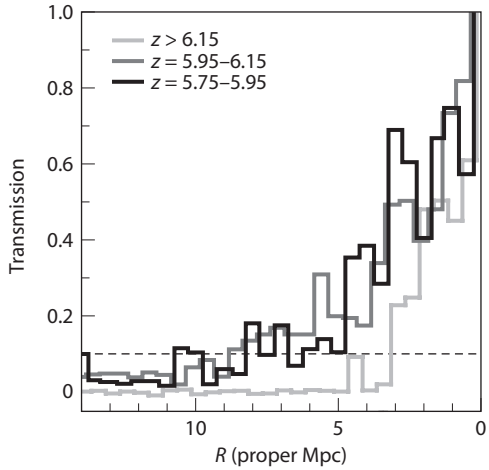


Figure 11.4 Average absorption profiles near the Lyman-α line for quasars in three different redshift bins. Note that the Lyman-α emission lines have been fitted and removed. The three redshift bins average over eight ($5.75 < z < 5.95$), nine ($5.95 < z < 6.15$), and four ($z > 6.15$) quasars. The horizontal dotted line marks 10% transmission, conventionally taken as the edge of the near-zone. Carilli, C. L., et al., *Astrophys. J.* **714**, L834 (2010). Reproduced with permission of the American Astronomical Society.

by /
AAS. /

is usually referred to as the *near-zone*, conventionally defined to be the point at which the transmission falls below 10%.

In the first case, the size of the H II region depends on the ionizing luminosity of the quasar (which can be estimated from the spectrum redward of Lyman-α), the age of the quasar t_Q, and the average neutral fraction before the quasar appeared, $\bar{x}_{\rm H\,I}$. The basic radiative transfer problem was solved in sections 9.1 and 9.8.2; for the purposes of a simple estimate, if recombinations can be neglected, the proper radius of the H II region is (cf. equation 9.3)[iv]

delete period between cf / PE

$$R_b \approx \frac{4.2}{\bar{x}_{\rm H\,I}^{1/3}} \left(\frac{\dot{N}_Q}{2\times 10^{57}\ {\rm s}^{-1}}\right)^{1/3} \left(\frac{t_Q}{10^7\ {\rm yr}}\right)^{1/3} \left(\frac{1+z}{7}\right)^{-1} {\rm Mpc}, \qquad (11.18)$$

where $\dot{N}_Q$ is the rate at which the quasar produces ionizing photons, and we have assumed that all the ionizing photons are absorbed but ignore secondary ionizations. Note that $R_b \propto (\dot{N}_Q t_Q/\bar{x}_{\rm H\,I})^{1/3}$ and varies relatively slowly with these parameters.

However, the absorption can become saturated well before this limit is reached. To estimate this point we suppose that the edge of the near-zone is where the optical depth rises above $\tau_{\rm lim}$. We adopt $\tau_{\rm lim} = 2.3$ as a fiducial value (see Figure 11.4). Assuming that the background ionization rate can be

[iv]Here we ignore relativistic effects in the expansion, which are important at early times. See equation (9.12) for a more accurate expression.

should read: no more than three to four times / PE

neglected (likely a good assumption at these very high redshifts), the transmission reaches this limiting value at a proper radius

$$R_{\rm lim} \approx 3.1\ {\rm Mpc} \left(\frac{\dot{N}_Q}{2\times10^{57}\ {\rm s}^{-1}}\right)^{1/2} \left(\frac{T}{2\times10^4\ {\rm K}}\right)^{0.38} \left(\frac{\tau_{\rm lim}}{2.3}\right)^{1/2} \times \left(\frac{3\alpha}{\alpha+3}\right)^{1/2} \left(\frac{1+z}{7}\right)^{-9/4}, \quad (11.19)$$

where the T dependence enters through the recombination coefficient for ionization equilibrium. Note that this limiting radius is independent of the neutral fraction of the material outside the ionized zone, and it is slightly more sensitive to the quasar luminosity, $R_{\rm lim} \propto \dot{N}_Q^{1/2}$.

There are two caveats on these size estimates. First, equation (11.19) can apply only if the quasar light has reached that distance. This requires

$$t_Q > 4.2\times10^6 \bar{x}_{\rm H\ I} \left(\frac{R_{\rm lim}}{3.1\ {\rm Mpc}}\right)^3 \left(\frac{\dot{N}_Q}{2\times10^{57}\ {\rm s}^{-1}}\right)^{-1} \left(\frac{1+z}{7}\right)^3 {\rm yr}. \quad (11.20)$$

(Adding recombinations and clumping will increase this scale by ~~three to four~~ no more than times.) Interestingly, this timescale is comparable to the canonical quasar lifetime $t_Q \sim 10^7$ yr in fully neutral gas, but for quasars positioned near the end of reionization (which are actually accessible to observations) it is very short.

Moreover, our expressions for R_b and $R_{\rm lim}$ implicitly ignore the possibility of LLSs or even denser regions in the IGM. If the quasar radiation encounters a highly overdense region that can maintain $\tau > 1$ in ionization equilibrium, the ionizing radiation will be highly attenuated at larger distances. Although these systems are likely to be rare near the quasar (where the radiation field is particularly strong), they are difficult to identify in the highly saturated forest spectra found during reionization, and they present an important systematic concern for measurements of *near-zones*.

We therefore expect most quasar near-zones to be limited by the proximity effect rather than the bubble size. If so, these zones can tell us little about the ionization state of the surrounding gas. In principle, this supposition can be tested by examining the luminosity dependence of the near-zone size, although the modest variation between the two models, and the large scatter intrinsic to any measurement in an inhomogeneous IGM, has made differentiating them difficult to date. Figure 11.5 shows the measured near-zone sizes for a number of quasars at $z > 5.75$. The left panel shows the trend with redshift (here all the near-zone sizes have been normalized to a common luminosity using the $R_b \propto \dot{N}_Q^{1/3}$ relation), while the right panel shows the dependence on absolute magnitude (with the mean trend over redshift removed).

In the right panel, the dotted curve shows $R_b \propto \dot{N}_Q^{1/3}$ (with arbitrary scaling); this is not a fit but is shown only for illustrative purposes. Clearly, the large scatter in the near-zone sizes, even after a simple redshift correction, makes it difficult to distinguish this behavior from that expected for the more classic proximity effect, $R_b \propto \dot{N}_Q^{1/2}$.

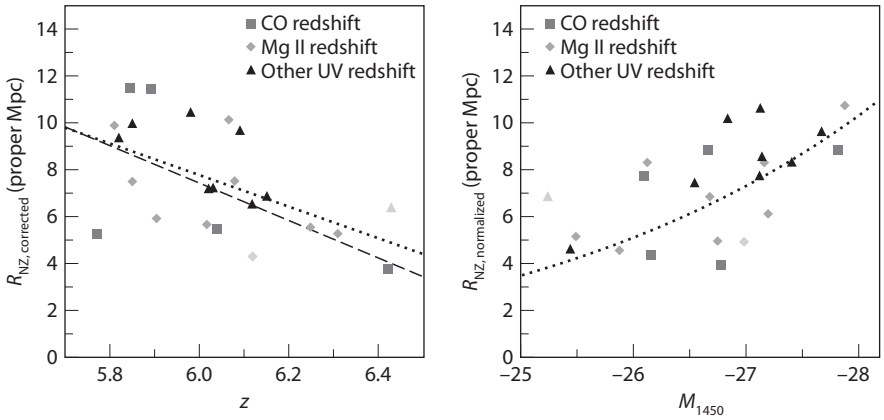


Figure 11.5 *Left:* Measured radii of near-zones in a set of high-z quasars; the symbols denote the method used to compute the quasar's redshift. All near-zone measurements have been scaled to a common quasar luminosity using the R_b relation in equation (11.18) to better illustrate the trend with redshift. Typical errors in the near-zone size are ~ 1 Mpc. The two lines are fits to the trend with redshift. *Right:* Dependence of the near-zone size on quasar absolute magnitude; all the data points have been scaled to a common redshift using the mean relation in the left-hand panel to better illustrate the behavior with luminosity. The dotted line shows $R_b \propto \dot{N}_Q^{1/3}$ with arbitrary scaling; note that it is not a fit but is merely meant to guide the eye. Carilli, C. L., et al., *Astrophys. J.* **714**, L834 (2010). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

Nevertheless, there is clearly a steady increase in the near-zone size as redshift declines. One possible interpretation is a decrease in $\bar{x}_{\rm H\,I}$ with cosmic time; the data require a decline by $\sim$10 over the range $z = 6.4$ to $z = 5.8$. However, presuming that $z \sim 6$ is the tail end of reionization, the proximity effect is more likely to fix the near-zone size. In that case, the trend with redshift is most likely attributable to a rapid increase in the background ionization rate (by a factor of >3), which can substantially boost the total ionization rate in the outskirts of the quasar's proximity zone.

Currently, the most challenging aspect of this measurement—other than finding these quasars in the first place—is determining the quasar's location. The only tools we have are the redshifts of the source's emission lines. Unfortunately, most quasars have strong internal motions and winds, which displace many of the emission lines from the systemic redshift of the host. The best choices are low-excitation lines (such as Mg II) or, even better, lines from the host galaxy itself. Any such lines in the optical or UV are overwhelmed by the quasar's own emission, so the most useful lines turn out to be those of CO, which are strong in these rapidly star-forming galaxies.

There is one additional, and very attractive, way to differentiate between R_b and $R_{\rm lim}$: by examining the absorption in higher Lyman-series lines. Because

$R_{\rm lim}$ depends on the maximum detectable optical depth $\tau_{\rm lim}^{1/2}$, it increases by the square root of the optical depth ratio between different lines; for Lyman-β, this means $R_{\rm lim}^{\beta} \approx 2.5 R_{\rm lim}^{\alpha}$. However, at the edge of the ionized bubble the neutral fraction presumably increases by an enormous amount over a very small distance, so both Lyman-α and Lyman-β should become optically thick at nearly the same radius. Unfortunately, this test is still sensitive to the large amount of scatter in the IGM density field (and in the lower-redshift Lyman-α forest that coincides with and hence obscures the Lyman-β measurement), so the current sample of <10 quasars cannot distinguish R_b from $R_{\rm lim}$—even though coincident Lyman-α and Lyman-β absorbers have been detected, it is not clear whether they are due to a large swath of neutral IGM gas or to a single absorber.[v] Simulations suggest that increasing the sample of such quasars by a factor of 3–5 could lead to useful constraints when $\bar{x}_{\rm H\,I} > 0.1$, the regime in which the finite bubble size starts to affect the Lyman-β near-zone size.

Another difficulty with near-zone measurements, just as with the classical proximity effect, is the biased region in which the quasar exists. Although the gas is only significantly overdense in a relatively small region immediately around the quasar, even modest overdensities in the dark matter can lead to substantial overdensities in the biased galaxy population. Moreover, the ionized bubble generated by these galaxies reaches much larger distances than the galaxy overdensity itself—even the tens of comoving megaparsecs typical of a bright quasar's near-zone. The easiest way to understand this is to think of the overdense region as a piece of a Universe with $\Omega_m > 1$: in that case structure formation proceeds faster, because of the increased gravity, and both the local ionized fraction and the ionized bubbles themselves grow faster as well. This implies that the ionized fraction measured from the quasar near-zone will be biased relative to the true average.

11.4 The Red Damping Wing

If photons generated near the Lyman-α line, but still redward of it, encounter nearly neutral gas with $\tau_\alpha > 10^5$, the broad Lyman-α absorption line can significantly affect their transfer through the IGM even though they remain relatively far from resonance. Considering only the regime in which $|\nu - \nu_\alpha| \gg \Lambda_\alpha$ (and neglecting the broadening introduced by the finite temperature of the IGM), we may ignore the second term in the denominator of equation (4.8). Assuming that the IGM has a uniform neutral fraction $\bar{x}_D$ at all points between the edge of a source's local ionized bubble (which we call z_b) and $z_{\rm reion}$, leads to an analytic result valid within the *red damping wing* of the Gunn-Peterson trough

delete comma / PE

[v]These kinds of identifications are further complicated by the damping wing absorption, which we examine next.

for the optical depth at an observed wavelength $\lambda_{\rm obs} = \lambda_\alpha(1+z)$:[5]

insert subscript Greek lowercase alpha / AA

$$\tau_D(z) = \tau_\alpha \bar{x}_D \left(\frac{\Lambda}{4\pi^2 \nu_\alpha}\right) \left(\frac{1+z_b}{1+z}\right)^{3/2} \left[I\left(\frac{1+z_b}{1+z}\right) - I\left(\frac{1+z_{\rm reion}}{1+z}\right)\right], \tag{11.21}$$

for $z > z_b$, where

$$I(x) \equiv \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\ln\left[\frac{1+x^{1/2}}{1-x^{1/2}}\right]. \tag{11.22}$$

Note that here we define z as the redshift at which the observed photon would have passed through Lyman-α; however, when $z > z_b$ this never happens. This expression is formally valid only far from line center, but that is usually acceptable, because the central optical depth is so large that the prediction there need not be precise. It also assumes $\Omega_m(z) \approx 1$, which is adequate at the high redshifts of interest, $z \gg 1$.

At wavelengths for which $|x - 1| \ll 1$, one can approximate the $I(x)$ factors with their asymptotic limits; in that case,

$$\tau_D(z) \approx \tau_\alpha \bar{x}_D \left(\frac{\Lambda}{4\pi^2 \nu_\alpha}\right) \frac{c(1+z)}{H(z)} \left(\frac{1}{R_{b1}} - \frac{1}{R_e}\right), \tag{11.23}$$

where R_{b1} is the comoving distance to the edge of the source's ionized bubble, and R_e is the comoving distance to the surface defining the "end" of reionization. As a rule of thumb, the damping wing optical depth approaches unity at a velocity offset of $\sim 1{,}500\,{\rm km\,s^{-1}}$, which corresponds to $\sim$1 proper Mpc at $z \sim 10$.

The exciting observational prospect is that within this red damping wing, the optical depth experienced by the photons is of the order of unity over a fairly wide range of redshifts: this means that the damping wing optical depth can be measured relatively easily, in contrast with the strongly saturated absorption at line center. Crudely, if we can therefore measure z_s and $\tau(z)$, we can obtain an estimate for the IGM neutral fraction.

Figure 11.6 illustrates the resulting absorption profiles for three choices of $\bar{x}_D = 0.9$, 0.5, and 0.1 (thin dashed, solid, and dotted curves, respectively); in all cases we take $z_{\rm reion} \ll z_b$. Here the abscissa measures the wavelength offset from the source redshift z_s; we take z_b, where neutral gas first appears, to be 5 comoving Mpc from the source. Note that, especially for the more neutral cases, the absorption extends to fairly large redshift offsets from line center: $z - z_s = 0.01(1+z_s)$ translates to an observed wavelength offset of $12(1+z_s)$ Å. The differences among these curves suggest that the profile of absorption *redward* of Lyman-α may be a powerful probe of the IGM ionization state.

The dot-dashed line in Figure 11.6 shows the absorption profile of a single absorbing cloud at a fixed location (i.e., a DLA), normalized to have the same transmission at z_s as the $\bar{x}_D = 0.1$ curve. Obviously, the IGM absorption profile is much gentler than that from a DLA, extending to much larger redshift offsets. Indeed, equation (11.23) shows that the optical depth scales as the inverse of the wavelength offset between the observed wavelength and λ_α at the

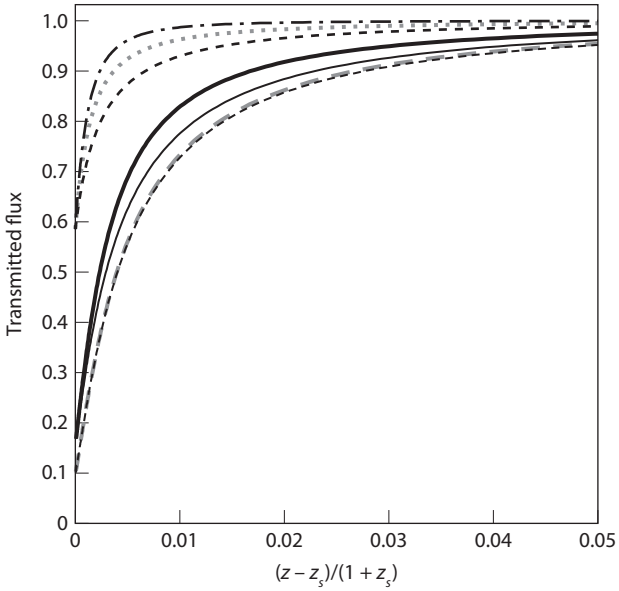


Figure 11.6 Damping wing absorption profiles as a function of fractional wavelength offset from the source (at redshift z_s). The thick curves show the absorption profiles for $\bar{x}_D = 0.9$, 0.5, and 0.1 assuming the "picket fence" model of absorption (with the dashed, solid, and dotted curves, respectively). The corresponding thin curves show the absorption profiles for uniformly ionized IGM normalized to the same transmission at z_s. Note that the two dashed curves overlap and are practically indistinguishable. The dot-dashed curve shows the profile of a DLA, normalized to the same transmission as the $\bar{x}_D = 0.1$ curves at z_s. Mesinger, A., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **385**, 1348 (2008). Copyright 2008 by the Royal Astronomical Society.

source redshift. In contrast, DLAs have $\tau \propto \Delta\lambda^{-2}$; the difference arises because a photon traveling through the IGM continues to redshift away from line center, so a photon at a given final (observed) wavelength therefore began its travels closer to line center and must have experienced a larger optical depth than one would expect had it remained at a constant frequency through the entire column. In practice, this may be a crucial discriminant between absorption intrinsic to a high-redshift source (taking the form of a DLA) and that from the IGM. For example, nearly all GRBs at lower redshifts have associated high-column absorbers.[6] The different absorption profiles are crucial for identifying the nature of the Lyman-α absorption.

Unfortunately, the simple toy model we have used so far does not accurately describe the IGM during reionization, and real damping wing absorption profiles are likely to be somewhere between these two limits. We have already seen that in most reionization scenarios the IGM has a two-phase structure, with seas of neutral gas surrounding bubbles of ionized matter. A typical line

of sight through the IGM will therefore pass through a "picket fence" of absorbers composed of alternating patches of nearly neutral and nearly ionized gas. The resulting absorption profiles, shown for a toy model by the thick lines in Figure 11.6, are steeper than those in a uniform IGM (unless the ionized bubbles are very rare) but shallower than for a DLA: essentially, the photon passes through a series of quasi-DLAs punctuated by clear, nonabsorbing H II regions. Nevertheless, because their frequency still changes as they travel, they experience more absorption than in a single cloud.

Obviously, this picture introduces some significant complications into interpreting the damping wing. The easiest way to see this is to consider a crude estimate for the average ionized fraction in a uniform IGM from inverting equation (11.23). Here we can estimate $\bar{x}_D$ from the absorption at a single wavelength, provided we assume a $1/\Delta\lambda$ profile to be accurate. (Note that we could also estimate z_b from the peak of the absorption line.) In this picket fence model, the true optical depth is a sum over those from all the neutral stretches of the IGM, or

insert subscript Greek lowercase alpha / AA

$$\tau(z) \approx \tau_\alpha \left(\frac{\Lambda}{4\pi^2 \nu_\alpha}\right) (1+z)^2 \sum_i \left(\frac{1}{z - z_{b,i}} - \frac{1}{z - z_{e,i}}\right), \tag{11.24}$$

where the ith neutral patch stretches between $z_{b,i}$ and $z_{e,i}$. If we naively equate this true expression to equation (11.23) and solve for $\bar{x}_D$, we find

$$\bar{x}_D \approx (z - z_{b,1}) \left\langle \sum_i \left(\frac{1}{z - z_{b,i}} - \frac{1}{z - z_{e,i}}\right) \right\rangle. \tag{11.25}$$

If we take a particularly simple model for the picket fence absorbers, in which the ionized and neutral patches have fixed lengths R_b and fR_b, where $f = (1 - Q_{\rm H\,II})/Q_{\rm H\,II}$ ensures the proper filling fraction of the bubbles, we can perform this sum and calculate the bias in our estimator $\bar{x}_D$:

$$\bar{x}_D = \frac{1}{2} \sum_{k=1}^{\infty} \left[\frac{1}{(k - 1/2) + (k-1)f} - \frac{1}{(k - 1/2) + kf}\right] \tag{11.26}$$

$$= \pi (1 - Q_{\rm H\,II}) \cot\left[\frac{\pi(1 - Q_{\rm H\,II})}{2}\right]. \tag{11.27}$$

This difference $\bar{x}_D - (1 - Q_{\rm H\,II})$ is always positive and peaks at ~ 0.3 when $Q_{\rm H\,II} = 0.5$, though the fractional bias continues to increase as $Q_{\rm H\,II} \to 1$. The actual amount of the bias of course depends on the particular model of reionization (and, in particular, the size distribution and clustering of the H II regions); more detailed simulations have comparable (though slightly smaller) bias. This means that the damping wing requires nontrivial modeling to interpret it properly in the context of reionization.

Even if this bias can be corrected, a second problem is that different lines of sight inevitably pass through different sets of ionized and neutral patches, so there can be large scatter in the absorption profiles even for a given $Q_{\rm H\,II}$ and bubble size distribution. This scatter becomes particularly important in the

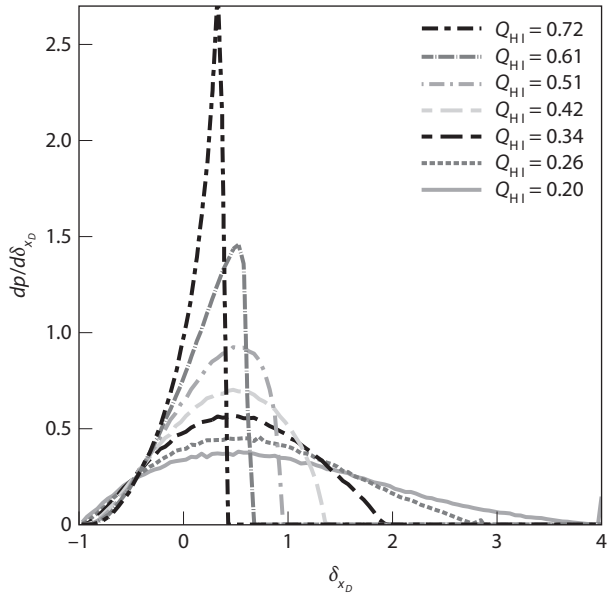


Figure 11.7 Probability distributions of the fractional bias in a simple damping wing estimate of the ionized bubble filling factor, $\delta_{x_D} \equiv \bar{x}_D/(1 - Q_{\rm H\,II}) - 1$. The different curves show different stages in reionization; all are computed with a seminumerical simulation of reionization. Note that the mean is always nonzero, and the distribution becomes both wider and more biased as reionization progresses. Mesinger, A., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **385**, 1348 (2008). Copyright 2008 by the Royal Astronomical Society.

late stages of reionization, because the damping wing optical depth is rather sensitive to the size of the first neutral patch.

Figure 11.7 illustrates these two problems in the context of a more realistic seminumerical model of reionization. The curves show the probability distribution of $\delta_{x_D} \equiv \bar{x}_D/(1 - Q_{\rm H\,II}) - 1$ for a variety of bubble-filling factors. Note that the means of these distributions are nonzero (implying a bias in the estimator), and the scatter increases dramatically in the later stages of reionization. This means that reliable estimates of the IGM properties will require a large number of lines of sight with measured damping wings.

Because the damping wing absorption profile must itself be measured at high signal to noise, damping wing constraints on reionization require very bright sources. The two most likely candidates are quasars and GRBs. The former have the advantage of lying inside large H II regions, which decreases the bias and scatter in the estimators; however, they often also have substantial Lyman-α lines with unknown intrinsic properties, which complicates the measurement of the damping profile.

Figure 11.8 illustrates some of the complexities of a damping wing measurement with a quasar at $z = 7.085$, ULAS J1120+0641. We show the spectrum

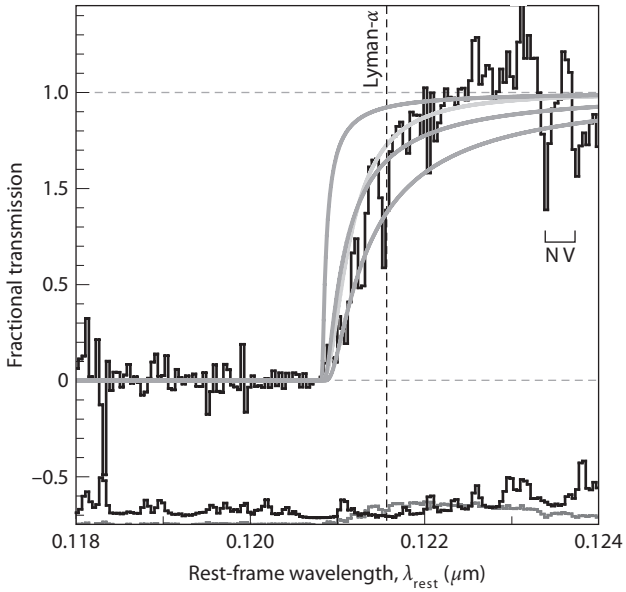


Figure 11.8 Rest-frame transmission profile of ULAS J1120+0641, in the region of the Lyman-α emission line, compared with several model absorption profiles. The transmission profile of ULAS J1120+0641, obtained by dividing the spectrum by a lower-redshift quasar composite spectrum, is shown as the binned curve. The random errors are plotted below the data (across all wavelengths). The other error curve shows the uncertainty in the Lyman-α equivalent width. Three of the four smooth curves in the upper panel show the expected absorption from an IGM damping wing with $\bar{x}_D = 1, 0.5, 0.1$ located a comoving distance $R_b = 17.8$ Mpc in front of the quasar (bottom, second from bottom, and top curves at the Lyman-α wavelength). The other curve (second from top) shows a DLA absorber with $N_{\rm H\,I} = 4 \times 10^{20}$ cm^{-2} located 21 Mpc in front of the quasar. Mortlock, D., et al., *Nature* **474**, 616 (2011). Copyright 2011 by Nature Publishing Group.

normalized to a composite constructed from lower-redshift quasars, which provides a surprisingly good fit to the data (though it appears to underestimate the Lyman-α line strength in this particular object). Provided the template is accurate, the binned curve in the figure therefore shows the measured transmission, which, as expected, declines rapidly slightly blueward of the Lyman-α line center; this is the near-zone discussed previously. This quasar has a very small near-zone, which indicates either a high–column density absorber along the line of sight, the presence of a substantially neutral IGM that the quasar still must ionize, or a very young age for the quasar. The smooth curves show the expected absorption for several IGM scenarios. The second curve from the top at the Lyman-α wavelength shows the absorption profile from a DLA 21 comoving Mpc in front of the source. The others show the absorption expected from a uniformly neutral IGM beginning 17.8 comoving Mpc in front of the quasar;

these take $\bar{x}_D = 0.1$, 0.5, and 1, from top to bottom. Of these, the DLA profile appears to provide the best fit; however, more sophisticated fits taking into account the inhomogeneous ionization structure of the IGM could also match the data.

GRBs have much simpler intrinsic spectra (nearly power law over this range), which makes extracting the damping wing easier. However, their host galaxies often have strong DLA absorbers, which interfere with the damping wing, and their position inside small galaxies makes the bias and scatter large. It is not clear which will eventually prove more useful, though in either case, constructing samples of many sources will be difficult. Also, in contrast with quasars, GRBs (and their faint host galaxies) have a negligible influence on the surrounding IGM, 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.

11.5 The Lyman-α Forest As a Probe of the Reionization Topology

lc / PE

Given the utility of the Lyman-α forest for understanding the ionization state of the IGM at low and moderate redshifts, the extension of these techniques to the cosmic dawn appears to provide an obvious test of the topology and nature of the reionization process. However, we have already seen that the Gunn-Peterson optical depth was large at that time, even in highly ionized gas. It is thus not obvious that we should expect a clear signature of the ionized bubbles.

Nevertheless, the nature of the transformation from a bubble-dominated IGM to the postreionization "web-dominated" IGM does offer some hope. Once the ionized bubbles became larger than the mean free path of the ionizing photons, the ionizing background saturated—even if the Universe were fully ionized, the metagalactic background would not increase. Thus, we can expect nearly as much transmission in bubbles that reached this saturation limit as in the postreionization IGM.

delete comma / AA

But there is one additional factor to consider: the damping wing from the neutral gas surrounding each ionized bubble. With the rule of thumb that $\tau_D < 1$ only at distances > 1 proper Mpc from fully neutral gas, this requires that ionized bubbles be at least a few megaparsecs large to allow for any transmission. Fortunately, in most reionization models this constraint is easily fulfilled, at least in the latter half of reionization (see Figure 9.5, for example).

Bubbles that allow transmission must be very large to build up a high enough ionizing background and avoid the damping wing, so they contain an enormous number of luminous sources. This in turn means that their ionizing background is fairly uniform (except at the edges of the bubble, but there the damping wing is large anyway). Thus, just as in the postreionization IGM, transmission will most likely come from highly underdense voids in which the neutral fraction is small. Equation (11.15) shows that $\tau_\alpha < 1$ requires $\delta < 0.1$–0.2.

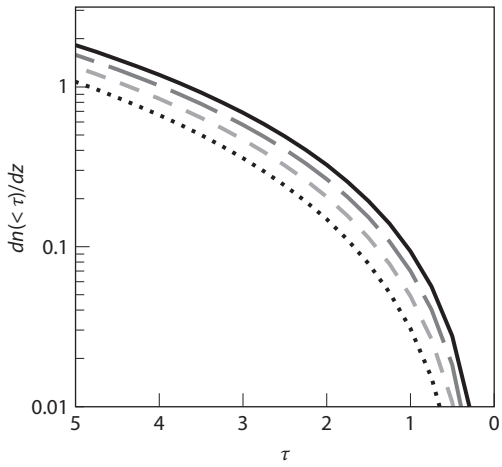


Figure 11.9 A model for the expected cumulative number of Lyman-α forest transmission features at $z = 6.1$ if the IGM has $Q_{\rm H\,II} = 0.9$, 0.85, 0.8, and 0.75 (solid, long-dashed, short-dashed, and dotted curves, respectively). The calculation uses the excursion set model for reionization (see §9.4) and an inhomogeneous IGM density distribution calibrated to simulations at lower redshifts. Furlanetto, S. R., et al., *Mon. Not. R. Astron. Soc.* **354**, 695 (2004). Copyright 2004 by the Royal Astronomical Society.

Such deep voids are very rare at high redshifts, because structure formation was still in its infancy— and, of course, such regions were very far from galaxy concentrations and so likely remained in neutral regions throughout nearly all of reionization.

Thus, we expect transmission spikes to be extremely rare (but not impossible) during reionization. With models for the H II region sizes, the emissivity of the galaxies driving reionization, and of the density distribution of the IGM, it is not difficult to estimate the possible abundance of transmission features. Figure 11.9 shows an optimistic example calculation for transmission at $z = 6.1$ (in the range probed by the highest-redshift known quasars). The curves show that observable transmission gaps with $\tau < 2.3$ occur only about once per total observed Lyman-α forest path length of $\Delta z \sim 3$.

In reality, transmission will be even more rare, because this simple calculation makes the optimistic assumption that photons travel to the edge of their bubble, without any limits from LLSs in the IGM. But even so, Figure 11.9 shows that they are sufficiently rare that precise quantitative constraints on reionization from any such transmission spikes will require much larger samples of quasars or GRBs than currently available. Drawing conclusions about reionization from the forest is instead very difficult. Indeed, some simulations of the reionization process show that the present data cannot even rule out reionization ending at $z < 6$, since some small pockets of neutral gas could remain, buried inside the long stretches of saturated absorption that were common at this time.[7]

Overall, the Lyman-α forest (especially together with absorption in Lyman-β and Lyman-γ) is best at constraining the very end of the reionization era, as discussed in §4.7, unless the red damping wing can be measured on its own.

11.6 Lyman-α Halos Around Distant Sources

lc / PE

11.6.1 The Scattering of Damping Wing Photons

As we have already discussed in the context of Lyman-α scattering within galaxies, Lyman-α line photons emitted by these galaxies are not destroyed but instead are absorbed and reemitted as they scatter. For scattering in the uniform IGM, this problem is particularly simple and illuminates more of the physics of the scattering process. For simplicity, we neglect the H II regions surrounding a galaxy here and imagine photons that begin on the red side of the line and scatter through the Hubble flow away from their source.

Owing to the Hubble expansion of the IGM around the source, the frequency of the photons is slightly shifted by the Doppler effect in each scattering event. As a result, the damping wing photons diffuse in frequency to the red side of the Lyman-α resonance. Eventually, when their net frequency redshift is sufficiently large, they escape and travel freely toward the observer (see Figure 11.10). As a result, the source creates a faint Lyman-α halo on the sky.[vi] These *Loeb-Rybicki* Lyman-α halos can be simply characterized by the frequency redshift relative to the line center, $v_\star = |v - v_\alpha|$, which is required to make the optical depth from the source equal to unity. At high redshifts, the leading term in equation (11.21) yields

$$v_\star = 8.85 \times 10^{12}\ \mathrm{Hz} \times \left(\frac{\Omega_b h}{0.05\sqrt{\Omega_m}}\right)\left(\frac{1+z_s}{10}\right)^{3/2} \tag{11.28}$$

as the frequency interval over which the damping wing affects the source spectrum. A frequency shift of $v_\star = 8.85 \times 10^{12}$ Hz relative to the line center corresponds to a fractional shift of $(v_\star/v_\alpha) = (v/c) = 3.6\times10^{-3}$, 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,

$$r_\star = 1.1\left(\frac{\Omega_b/0.05}{\Omega_m/0.3}\right)\ \mathrm{Mpc}. \tag{11.29}$$

Typically, the observable Lyman-α halo of a source at $z_s \sim 10$ 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 11.10 and 11.11. The scattered photons are highly polarized, with the polarization direction determined by the orientation

[vi] The photons that begin blueward of Lyman-α and get absorbed in the Gunn-Peterson trough are also reemitted 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 than do 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.

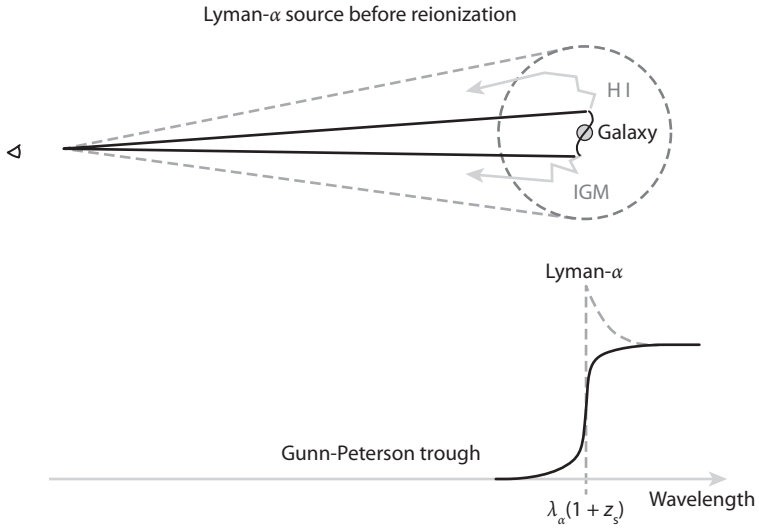


Figure 11.10 Halo of scattered Lyman-α line photons from a galaxy embedded in the neutral IGM prior to reionization (also called a *Loeb-Rybicki halo*); see *Color Plate 24* for a color version of this figure. The line photons diffuse in frequency owing the Hubble expansion of the surrounding medium 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 $\sim$1,000 km s^{-1} relative to other lines (such as Hα) emitted by the galaxy.

to / PE

of the observer relative to the last scattering atom. Thus, the shape of the halo would be different if viewed through a polarized filter.[8]

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 gas distribution and the velocity field of the neutral IGM before the epoch of reionization. The Lyman-α sources serve as lamp posts that illuminate the surrounding H I fog. However, owing to their low surface brightness, the detection of Lyman-α halos through a narrowband filter is much more challenging than direct observation of their sources. Moreover, the velocity fields around these galaxies may be complicated by winds and infall, which would affect the line brightness and profile in similar ways to those discussed in §11.1.1.

11.6.2 Lyman-α Blobs

A particularly interesting example of Lyman-α line emission in the interface between galaxies and the IGM are the so-called "Lyman-α blobs" (LABs) originally discovered in narrowband images at moderate redshifts ($z\sim3$).[9] So far, several tens of LABs have been found in the redshift range $z \sim 2$–7, making

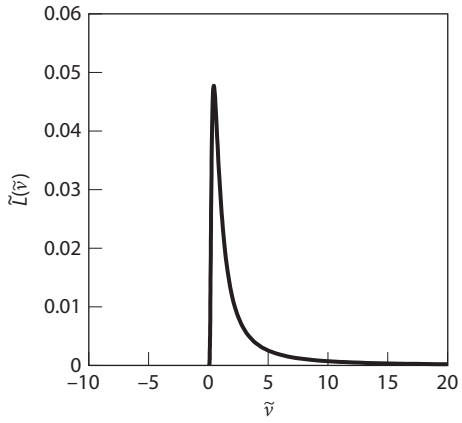


Figure 11.11 Monochromatic photon luminosity of a Lyman-α halo as a function of normalized frequency shift from the Lyman-α resonance, $\tilde{v} \equiv (v_\alpha - v)/v_\star$. Note that only the photons inside the red damping wing scatter in this compact halo; those on the blue side of Lyman-α scatter at much larger distances. The observed spectral flux of photons $F(v)$ (in photons cm^{-2} s^{-1} Hz^{-1}) from the entire Lyman-α halo is $F(v) = (\tilde{L}(\tilde{v})/4\pi\, d_L^2)(\dot{N}_\alpha/v_\star)(1+z_s)^2$, where $\dot{N}_\alpha$ is the production rate of Lyman-α photons by the source (in photons s^{-1}), $v = \tilde{v}v_\star/(1+z_s)$, and d_L is the luminosity distance to the source. Loeb, A., & Rybicki, G. B., *Astrophys. J.* **524**, 527 (1999). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

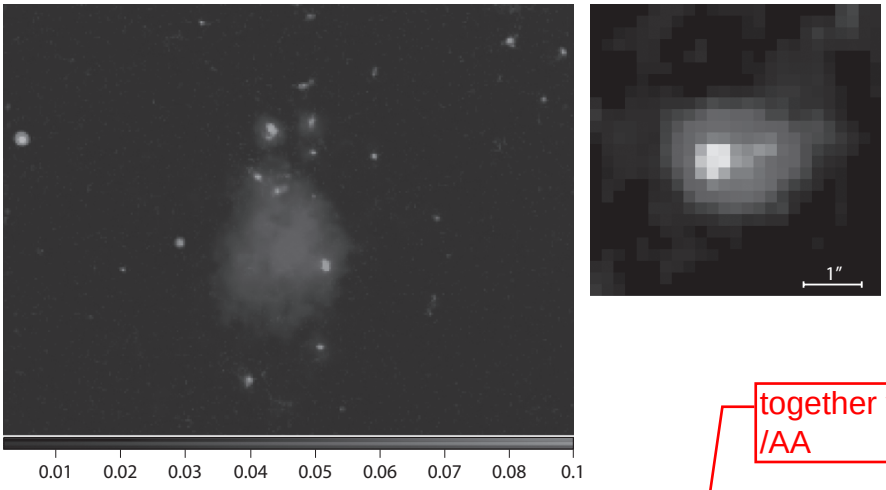
by / AAS. /

them much more common than initially expected.[10] These blobs have a range of properties, but all are characterized by significantly extended Lyman-α line emission (ranging in size from $\sim$10 kpc "halos" around star-forming galaxies to >150 kpc giants with no obvious central galaxy in the rest-frame UV). Some appear to be diffuse elliptical objects, while others are much more filamentary. The brighter objects, with line luminosities $L > 10^{44}$ erg s^{-1}, are extraordinarily powerful, corresponding to star formation rates $> 50\, M_\odot$ yr^{-1} if their energy source is attributed to obscured star formation. The lines can be quite broad but do not show any unusual features like double-peaked profiles. Two example objects are shown in Figure 11.12.

insert space before 1 / PE

Bright LABs are typically located near massive galaxies that reside in dense regions of the Universe. Multiwavelength studies of LABs reveal a clear association of the brighter blobs with submillimeter and IR sources that form stars at exceptional rates of $\sim 10^3\, M_\odot$ yr^{-1}, or with obscured active galactic nuclei (in fact, strong Lyman-α emission has been known for many years to surround some high-redshift radio galaxies).[11] However, other blobs have been found that are not associated with any source powerful enough to explain the observed Lyman-α luminosities.[12]

The origin of LABs is still unclear. Some models relate LABs to cooling radiation from gas assembling into the cores of galaxies.[13] Other models invoke photoionization of cold ($T \sim 10^4$ K), dense, spatially extended gas by an obscured



together with the /AA

Figure 11.12 *Left:* A false-color image of a Lyman-α blob (LAB) at redshift $z = 2.656$. The hydrogen Lyman-α emission is shown ~~in blue, and~~ images in the optical V-band and the near-IR 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. Prescott, M., & Dey, A. (2010). *Right:* A false-color image of a LAB at redshift $z = 6.6$, obtained from a combination of images at different IR wavelengths. (See *Color Plate 25* for a color version of this figure.) Ouchi, M., et al., *Astrophys. J.* **696**, 1164 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

move sentence in parenthesis after end of above sentence. /AA

by / AAS. /

quasar[14] or extended X-ray emission;[15] the compression of ambient gas by superwinds to a dense Lyman-α emitting shell;[16] or star formation triggered by relativistic jets from AGN.[17] The latest models relate LABs to filamentary flows of cold ($\sim 10^4$ K) gas into galaxies, which are generically found in numerical simulations of galaxy formation (see §8.2).[18] These cold flows contain ~5–15% of the total gas content in halos as massive as $M_{\rm halo} \sim 10^{12}$–$10^{13}\,M_\odot$.[19]

Although these objects have been observed in detail so far only at low redshifts, similar mechanisms offer the prospect of learning not only about star formation inside high-redshift galaxies and the gross properties of the IGM but also about the detailed structure of the gas accreting onto, or flowing out of, young galaxies. Lyman-α studies may therefore ultimately hold the key to understanding the initial stages of galaxy formation and growth.

11.6.3 Lyman-α Emission From the Intergalactic Medium

lc / PE

One additional source of diffuse Lyman-α emission may be important in the prereionization era: photons generated by IGM gas.[20] In principle, they may allow us to image the ionized phase as it filled the Universe.

Conceptually, there are two possible sources for such emission. First, in any portion of the ionized medium, the gas will continually recombine, generating Lyman-α photons at roughly two-thirds of this recombination rate. These photons will then propagate toward the observer. If we characterize the clumpiness of the volume by $C \equiv \langle n_e^2 \rangle / \langle n_e \rangle^2$, this Lyman-$\alpha$ photon production rate, per unit proper volume, is

$$\dot{n}_{\alpha,\mathrm{IGM}} = \frac{2}{3} \alpha_B \langle n_e \rangle \langle n_p \rangle C, \tag{11.30}$$

italics / AA

where we conservatively adopt the case-B recombination rate, assuming that the ionizing photons generated by recombinations directly to the ground state escape to fully neutral gas at the edge of the ionized region. A region of volume V will therefore produce a number flux of Lyman-α photons $\sim \dot{n}_{\alpha,\mathrm{IGM}} V / 4\pi\, d_L^2$. Assuming that the region expands at the usual Hubble flow rate, the relevant column density over a given observing band is simply determined by the radial distance corresponding to the band, so the surface brightness per unit wavelength of the emission[vii] is

$$\frac{d\Phi}{d\lambda_{\mathrm{obs}}} \approx \frac{1}{4\pi} \frac{\dot{n}_{\alpha,\mathrm{IGM}}}{(1+z)^3} \frac{dr_p}{dz} \frac{dz}{d\lambda_{\mathrm{obs}}} \tag{11.31}$$

$$\approx 0.032\, C \left(\frac{1+z}{10} \right)^{1/2} \text{ photons cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}\ \text{Å}^{-1}, \tag{11.32}$$

where the factor $(1+z)^{-3}$ is the usual cosmological surface dimming, $(d_A/d_L)^2$, modified for our definition in terms of photon number flux rather than energy flux.

In principle, a spectrum along any given line of sight will thus show an enhancement in the background at $\lambda_\alpha(1 + z_r)$, where z_r is the location of the reionization surface, followed by an alternating pattern of high and low intensity as the line of sight passes through the patchwork of H II regions. However, unless the IGM is extremely clumpy, the surface brightness provided by equation (11.32) is very small and unlikely to be detectable in any reasonable observation for the foreseeable future.

Fortunately, there is a second effect that can substantially increase the photon flux emerging from the hot "skin" of an ionization front propagating into the nearly neutral IGM: collisional excitation. Hydrogen line cooling is by far the most important cooling mechanism for gas of primordial composition and $T \sim 10^4$–10^5 K, and a substantial fraction of that line emission is deposited in the Lyman-α line. The emissivity in this case is

$$\dot{n}_{\alpha,\mathrm{CE}} = n_e n_{\mathrm{H\ I}} q_{\mathrm{Ly}\alpha}^{\mathrm{tot}}(T), \tag{11.33}$$

where $q_{\mathrm{Ly}\alpha}^{\mathrm{tot}}$ is the rate at which collisions generate Lyman-α photons (including those generated through cascades from higher electronic levels). This last function increases rapidly near $T \sim 10^4$ K and peaks at $T \sim 4 \times 10^5$ K; however, at

[vii] The surface brightness is the flux per solid angle from an object, or the flux divided by A_s/d_A^2, where A_s is the proper area of the object.

those high temperatures the gas is nearly entirely ionized, so $n_{\rm H\ I}$ is small, and the Lyman-α emission rate decreases again.

Put another way, the *total* amount of Lyman-α emission thus depends on the thermal energy imparted to each parcel of gas during the photoionization process: $\dot{n}_{\alpha,\rm CE} t_{\rm cool} \sim nk(T_i - T_f)$, where T_i is the initial post-ionization temperature, and T_f is the final value. The physics of line excitation then determines the relevant cooling time as well as a factor of the order of unity that regulates the fraction of the thermal energy released as Lyman-α photons. The strongest emission will therefore surround spectrally hard sources, which heat the gas to high temperatures (see §9.9).

Because $q^{\rm tot}_{\rm Ly\alpha}$ depends sensitively on temperature, and because $n_{\rm H\ I}$ changes rapidly within the ionization front, a detailed estimate of the surface brightness requires a radiative transfer simulation. A simple estimate, informed by such simulations, is simply to integrate the specific intensity through the ionization front, so that

$$I_{\rm Ly\alpha} \approx \frac{1}{\pi} \int n_{\alpha,\rm CE}\, dr \tag{11.34}$$

$$\sim \frac{f}{\pi} C A n_p^2 \left(\frac{1}{\sigma_{\rm H\ I,\langle \nu \rangle} n_p} \right) q^{\rm tot}_{\rm Ly\alpha}(\langle T \rangle), \tag{11.35}$$

where A is a geometric factor of the order of unity describing the run of $x_{\rm H\ I}$ through the ionized front, and we have evaluated the collisional excitation rate at an effective temperature $\langle T \rangle$. The factor $1/\pi$ accounts for the fraction of photons directed toward the observer. We have also introduced the clumping factor C to describe the enhancement in the collision rate inside a clumpy medium. Finally, we have characterized the thickness of the ionization front by the mean free path of a photon at a mean energy $\langle \nu \rangle$, and f is the fraction of the Lyman-α photons that escape to the observer.

The corresponding surface brightness (integrated across the entire ionization front) is then $\Phi \sim I_{\rm Ly\alpha}/(1+z)^3$, which is *independent* of redshift because the decreased thickness of the ionization front compensates for surface brightness dimming. With $\langle T \rangle = 3 \times 10^4$ K and $\langle \nu \rangle = 3 \nu_{\rm H\ I}$,[21]

$$\Phi \sim 10\, C \left(\frac{f}{0.5} \frac{A}{1/4} \right) \text{ photons cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}. \tag{11.36}$$

The fiducial choice for A simply takes $x_{\rm H\ I}(1 - x_{\rm H\ I}) \sim 1/4$ in the primary emission zone.

Even with reasonably optimistic clumping factors of $C \sim 5$, this is still a rather low surface brightness, and its detection presents a substantial challenge for the largest telescopes available. If a signal can be detected, another challenge remains in distinguishing IGM emission from the integrated background generated inside the ISM of unresolved high-redshift galaxies (see §13.2.1). However, searches around known bright quasars with relatively large H II regions allow one to integrate over substantial areas on the sky and search for a "coherent" signal representing the entire outer boundary of the ionized bubble.

11.7 Lyman-α Emitters During the Reionization Era

We now return to discuss the properties of more normal galaxies that have Lyman-α lines, commonly referred to as Lyman-α emitters or LAEs. We saw in §10.2.1 that this strong emission line provides a convenient marker for young star-forming galaxies, and one of the most efficient ways to find distant galaxies is with narrowband searches that identify sources with strong emission lines in a narrow redshift range.

We saw ~~seen~~ in §11.1.1 that the intrinsic properties of the Lyman-α line depend on a host of complex factors. However, we also found in §11.3.1 and §11.4 that resonant absorption in the ionized IGM and much stronger absorption from neutral gas—even from the damping wing for photons that do not pass through resonance—can also strongly affect the line (both in its amplitude and profile); this latter factor is denoted $T_{\rm IGM}^{\rm Ly\alpha}$ in equation (11.2). These effects make the Lyman-α emission lines of galaxies an interesting and potentially powerful probe of IGM properties. However, we must always bear in mind the complexity of the intrinsic line as an important source of systematic confusion for such a probe.

Figure 11.13 shows how this IGM reprocessing can dramatically alter the observed line intensity and profile; the top panel shows the lines, while the bottom panel shows the corresponding optical depth profiles. In the top panel, the upper dotted curve shows the assumed intrinsic line, which we place at $z = 10$ and take as a Gaussian with width 27 km s^{-1} (these are arbitrary choices chosen for illustrative purposes). The other curves show the effects of IGM reprocessing, including both the damping wing from fully neutral gas beginning a distance R_b from the line source (with R_b decreasing from top to bottom) and resonant scattering from the ionized medium within (except for the lower dotted curve). The optical depths providing this absorption are shown in the bottom panel: the nearly horizontal lines are the damping wing optical depths (with R_b ~~increasing~~ from bottom to top), while the dotted curve shows the resonant value.

Note that the resonant absorption is large everywhere blueward of line center but is modest or negligible on the red side. This is a rather generic result (here we have included only the ionization from the galaxy itself, which dominates on the relevant scales, so the ionization structure on large scales is negligible); in general, we expect LAEs at $z > 5$ to have asymmetric line profiles, with the blue side cut off by resonant IGM absorption.[viii]

However, the damping wing absorption that affects the red side (as well as the blue side) depends sensitively on the large-scale environment and, in particular, the displacement from the source to the nearest neutral gas. As described previously, a bubble with $R_b = 1$ *proper* Mpc produces $\tau_D \approx 1$; in fact, this rule of thumb works reasonably well throughout the relevant high-z regime.

[viii] In principle, this could be avoided if the line center was displaced redward of the galaxy's redshift, owing to reprocessing through a wind (see §11.1.1). However, nearly all observed LAEs (and Lyman-α emission lines from other moderate- and high-redshift galaxies) are asymmetric, even if they have large wind velocities.

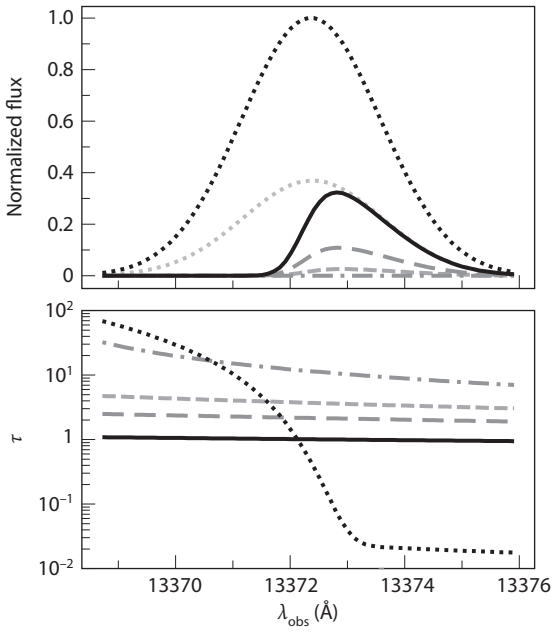


Figure 11.13 *Top:* Example line profiles for a galaxy at $z = 10$. The upper dotted curve shows the intrinsic line profile, assumed to be a Gaussian with standard deviation 27 km s^{-1}. The solid, long-dashed, and short-dashed curves show the observed line after reprocessing through the IGM; they place the galaxy in ionized bubbles with radii $R_b = 10$, 5, and 3 comoving Mpc, respectively. The lower dotted curve shows the line if we neglect resonant absorption within the ionized bubble, assuming $R_b = 10$ Mpc. *Bottom:* The dotted line shows the resonant absorption from the ionized bubble. The solid, long-dashed, short-dashed, and dot-dashed curves show the damping wing optical depth for $R_b = 10$, 5, 3, and 1 Mpc, respectively. Furlanetto, S. R., et al., *Mon. Not. R. Astron. Soc.* **354**, 695 (2004). Copyright 2004 by the Royal Astronomical Society.

We therefore expect that as we penetrate farther back into the reionization era, with the bubbles growing smaller and smaller, more and more of their Lyman-α lines will be extinguished by the neutral gas. In the remainder of this section we explore the consequences of this naive expectation for LAE surveys during reionization.

11.7.1 Galaxies within Ionized Bubbles

To understand the interplay between the damping wing and galaxy populations, we must first understand how galaxies populate the H II regions that surround them. Fortunately, because we can use the same method—the excursion set formalism—to compute the halo and ionized bubble abundances, this is a relatively easy task.

Consider an ionized bubble with mass m_b and a mean overdensity δ_b; according to the model in §9.4, this overdensity is exactly that required for a collapse fraction large enough to produce one ionizing photon per hydrogen atom inside the bubble, so $\delta_b = B(m_b)$.[ix] We wish to know the abundance of galaxies as a function of mass m within this ionized bubble, $n(m|m_b)$.

In the excursion set picture (see §3.4.2), this is simply proportional to the fraction of random walks that begin at (m_b, δ_b) and end at $(m, \delta_{\rm crit})$, where $\delta_{\rm crit}$ is the critical linearized overdensity for halo collapse (which is a function of m in, for example, the Sheth-Tormen model). But this problem is actually identical with the "extended Press-Schechter" problem, in which we calculated the progenitors of a given halo at an earlier redshift: the only difference is that here our "descendant halo" is a bubble and we work at the same redshift—which is possible because the criterion for an ionized bubble requires a lower overdensity than halo collapse itself.

Thus we can immediately write

$$n(m|m_b) = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{m^2}\left|\frac{d\ln\sigma}{d\ln m}\right|\frac{\sigma^2[\delta_{\rm crit}(z) - B(m_b)]}{(\sigma^2 - \sigma_b^2)^{3/2}}\exp\left\{-\frac{[\delta_{\rm crit}(z) - B(m_b)]^2}{2[\sigma^2 - \sigma_b^2]}\right\}, \tag{11.37}$$

where $\sigma^2 = \sigma^2(m)$, and $\sigma_b^2 = \sigma^2(m_b)$.

We can also perform the reverse calculation (analogous to the distribution of halo descendants) to compute the probability $p_b(m_b|m)$ that a halo of mass m is part of a bubble of mass m_b. Figure 11.14 shows the results of this calculation for a small halo ($m_h = 10^9\,M_\odot$) and a large one ($m_h = 10^{11}\,M_\odot$). The different curves in each panel correspond to a sequence of ionized fractions in a model of reionization. Unsurprisingly, the median bubble size increases as reionization progresses (because all bubbles grow with time), but note that it also strongly depends on the halo mass: large galaxies are far more likely to reside in large bubbles than are average galaxies. This is just another manifestation of the increasing bias of galaxies with their mass.

11.7.2 LAE Number Counts during Reionization

Next, let us imagine performing a sequence of narrowband LAE searches at progressively larger redshifts. We expect that once the typical bubble size falls below ~ 1 proper Mpc, the IGM damping wing will start to extinguish the Lyman-α emission lines even if the galaxies still exist. We might therefore imagine a simple counting exercise as a test for reionization, aiming to see a decline in the abundance of LAEs.

Of course, there are many other reasons why the LAE density may decline—most obviously, the halo mass function changes rapidly with z at these early times, so the galaxy abundance most likely does as well. Ideally, one would therefore calibrate the experiment to a broadband galaxy survey that is not subject to the same selection effects—a precipitous decline in LAE abundance with

[ix] Here, for simplicity, we ignore recombinations in the calculation.

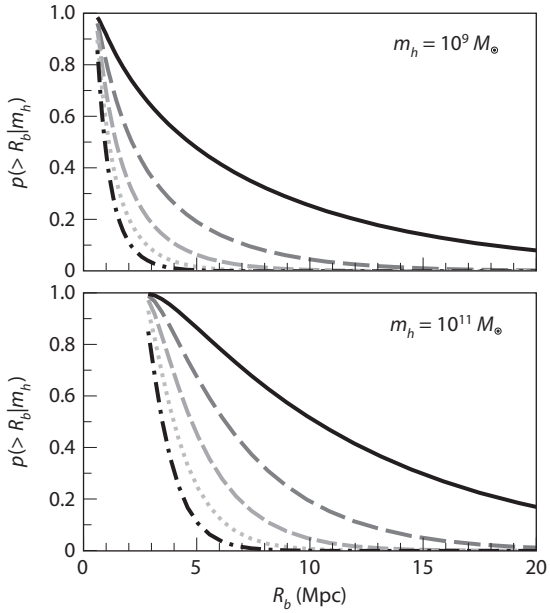


Figure 11.14 Probability that halos with $m_h = 10^9$ and $10^{11}\ M_\odot$reside in ionized bubbles larger than a given radius R_b. Here we use the excursion set model of reionization with $\zeta = 40$; the bubble sizes are relatively independent of this choice, for a fixed $Q_{\rm H\ II}$, but the halo populations themselves are highly redshift dependent. In each panel, the curves correspond to $z = 12$ ($Q_{\rm H\ II} = 0.74$, solid), $z = 13$ ($Q_{\rm H\ II} = 0.48$, long-dashed), $z = 14$ ($Q_{\rm H\ II} = 0.3$, short-dashed), $z = 15$ ($Q_{\rm H\ II} = 0.19$, dotted), $z = 16$ ($Q_{\rm H\ II} = 0.11$, dot-dashed) Furlanetto, S. R., et al., *Mon. Not. R. Astron. Soc.* **354**, 695 (2004). Copyright 2004 by the Royal Astronomical Society.

only a gentte decline in the overall galaxy density would be good evidence for IGM absorption. Similarly, one could also imagine identifying a galaxy sample using photometric techniques (as in §10.2.2) and then following them up spectroscopically to determine how their Lyman-α lines evolve. Note, however, that the complicated physics of Lyman-α generation and transfer within galaxies always leaves some room for doubt, since such a decline could also be attributed to the evolving IMF of stars or changes in their dust content.

Nevertheless, this simple test is very attractive. We can use the excursion set formalism described in §11.7.1 to estimate how the abundance would decline. We ignore the effects of resonant absorption (since they depend on the local environment of the galaxy and hence are unlikely to evolve rapidly during reionization) but include the damping wing absorption from neutral IGM gas. Let us suppose that the survey is sensitive to all sources with $L > L_{\rm min}$. If we then take $L \propto m$ for simplicity, a galaxy halo of mass m will be detected only if the damping wing has $\tau_D < \ln(m/m_{\rm min})$, where $L(m_{\rm min}) = L_{\rm min}$. Then, the number

density of observable galaxies is

$$n(>L) = \int dm_b\, n_b(m_b) V_b \int_{m_D}^{\infty} dm\, n(m|m_b), \tag{11.38}$$

where m_D is the minimum halo mass that remains observable inside a bubble of mass m_b and volume V_b. Note that m_D *decreases* with m_b, since larger bubbles cause less damping wing absorption. Of course, in reality, τ_D is a function not only of bubble size but of a galaxy's position within the bubble: those at the edge always experience strong absorption.

This simple model is in good agreement with more detailed calculations using simulations of reionization (either full-scale or seminumerical). Figure 11.15 shows the luminosity function at several different neutral fractions (including fully ionized, top curve) measured in a seminumerical simulation. Clearly, damping wing absorption from the neutral gas can have an enormous effect on the observed abundance of galaxies in these surveys.

The detailed calculation reveals two interesting effects. First, the fractional decline is relatively modest (no more than a factor ~2) until $Q_{\rm H\,II} < 0.5$; beyond that point the abundance declines steeply, because the ionized bubbles have characteristic sizes ~ 10 comoving Mpc, or ~ 1 proper Mpc, when $Q_{\rm H\,II} \sim 0.5$. Larger bubbles, late in reionization, have $\tau_D < 1$ and so have only a small effect on the observed abundance.

needs space / PE

The second factor is visible in the bottom panel of Figure 11.15: evidently, the fractional decline in LAE abundance is nearly independent of halo mass (or intrinsic luminosity). This occurs because the distribution of τ_D is quite broad (roughly lognormal), owing not only to the range of halo sizes but also to the distribution of galaxies within each bubble. For faint galaxies, which roughly follow a power-law intrinsic distribution, the convolution of these two effects preserves the power law. At the bright end, where the intrinsic luminosity function declines exponentially, the breadth of the τ_D distribution masks the change in slope.

11.7.3 LAE Clustering during Reionization

The fact that galaxies within large ionized bubbles remain (relatively) unattenuated while those inside small bubbles are extinguished by the damping wing suggests that not only did the mean number density of LAEs evolve throughout reionization, but their spatial distribution evolved as well. Figure 11.16 shows this explicitly. Each panel shows a slice through a seminumerical simulation of reionization; here we fix $z = 9$ and vary the ionized fraction across the panels (from fully ionized at left to $Q_{\rm H\,I} \equiv 1 - Q_{\rm H\,II} = 0.77$ at right). Each white dot corresponds to a galaxy with an observable Lyman-α line, assuming the same model as the last section for its luminosity function. The overall trend is clear: galaxies that are relatively isolated in the leftmost panel disappear first, while those that are part of a strong overdensity (near the bottom center of the image) remain visible even to large neutral fractions.

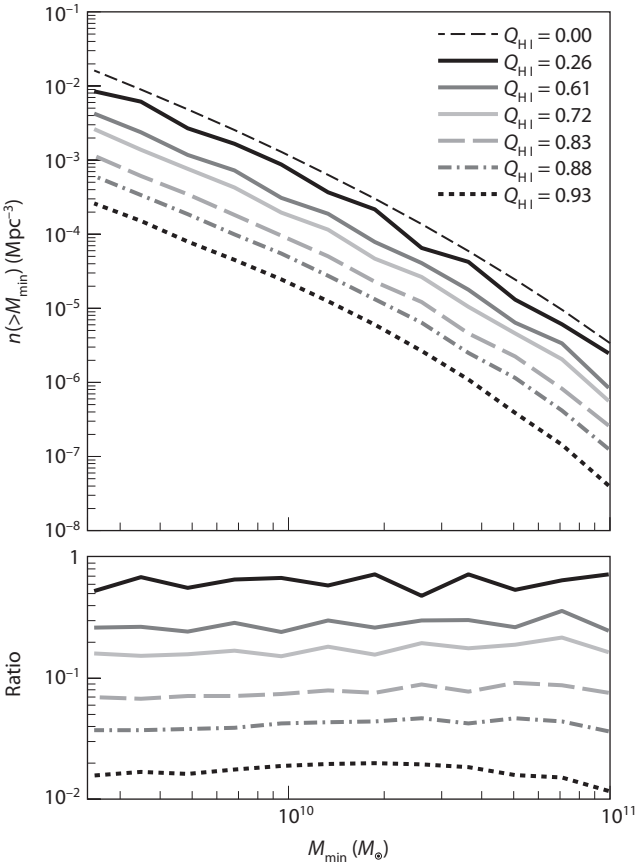


Figure 11.15 Luminosity function of LAEs at $z = 9$ in a seminumerical simulation of reionization, as a function of the mean neutral fraction $Q_{\mathrm{H\,I}} = 1 - Q_{\mathrm{H\,II}}$. The sequence of curves from top to bottom is from small to large $Q_{\mathrm{H\,I}}$. The bottom panel shows the ratio of the curves to that in a fully ionized Universe. Mesinger, A., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **386**, 1990 (2008). Copyright 2008 by the Royal Astronomical Society.

The best way to describe this phenomenon quantitatively is through the clustering of the galaxies. A simple toy model illustrates how it enhances the apparent clustering on small scales (relative to galaxies observed in the continuum, for example); see Figure 11.17. Suppose that galaxies with number density $\bar{n}$ are distributed randomly throughout the universe, but we can observe only those with at least one neighbor within a sphere of volume $V \ll \bar{n}^{-1}$. In other words, the ionized bubbles surrounding such clumped sources overlap, creating a large enough common bubble to hold the damping wing at bay. Assuming a Poisson distribution, the number density of observed objects would be

$$n_{\mathrm{obs}} = \bar{n}(1 - e^{-\bar{n}V}). \tag{11.39}$$

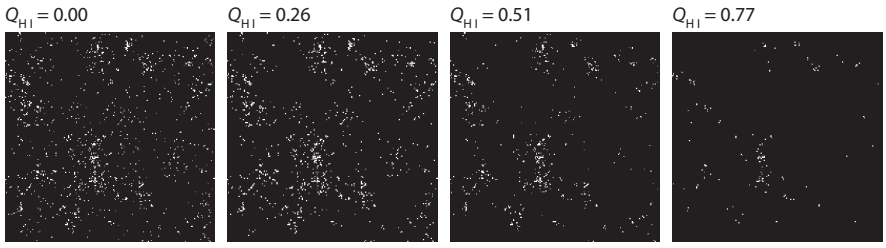


Figure 11.16 Maps of visible LAEs at $z = 9$ in a seminumerical simulation, assuming $Q_{\rm H\,I} \approx 0$, 0.26, 0.51, 0.77, from left to right. All slices are 250 Mpc on a side and 20 Mpc deep. We assume that all halos with observed luminosities greater than that corresponding to an unattenuated galaxy with $M_h > 1.67 \times 10^{10}\, M_\odot$ are visible and that $L \propto M_h$. Mesinger, A., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **386**, 1990 (2008). Copyright 2008 by the Royal Astronomical Society.

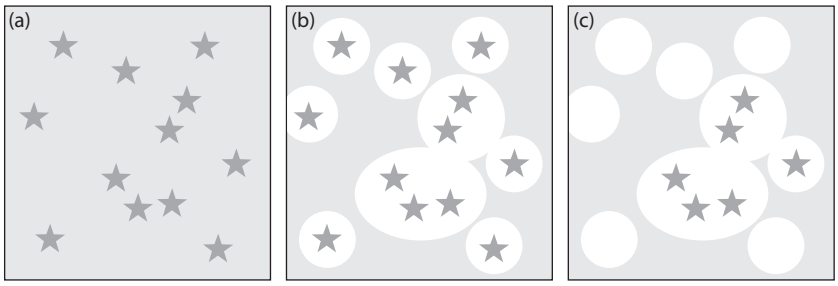


Figure 11.17 Toy model for the excess clustering induced in a Lyman-α selected galaxy population during reionization. *(a)* First, we assume a randomly distributed galaxy population. A continuum survey, sensitive to all these galaxies, would detect no clustering signal. *(b)* The H II regions generated by these galaxies: note that when galaxies are close together, their H II regions overlap, producing long stretches of nearly ionized gas through which Lyman-α photons can easily propagate. *(c)* A LAE survey would see galaxies only inside these large H II regions, all of which have neighbors (by construction), so the *apparent* clustering would be large even though the underlying population is randomly distributed.

As usual, the correlation function of the observed sample is defined through the total probability of finding two galaxies in volumes δV_1 and δV_2,

$$\delta P = n_{\rm obs}^2\,(1+\xi)\,\delta V_1\,\delta V_2. \tag{11.40}$$

However, we know that every observed galaxy has a neighbor within V; thus,

$$\delta P = n_{\rm obs}\,\delta V_1\,(\delta V_2/V) \tag{11.41}$$

for small separations (where the factor $\delta V_2/V$ assumes the neighbor to be randomly located within V). Thus,

$$\xi = 1/(n_{\rm obs}V) - 1 \tag{11.42}$$

on such scales: even though the underlying distribution is random, the selection criterion induces clustering. Note that it can be extremely large if $V \ll n_{\rm obs}^{-1}$.

On large scales, the modulation takes a different form. An observed galaxy resides in a large bubble, corresponding to an overdense region. Because of the bias of the underlying dark matter field, that overdense region will tend to lie near other overdense regions—and hence other large bubbles. Thus, we are more likely to see galaxies near the original object than in an average slice of the universe. Because we do not see similar galaxies in small (less-biased) bubbles, the large-scale bias is generically larger than that intrinsic to the galaxies.

Because these two effects have different amplitudes, the bubbles introduce a scale-dependent bias to the correlation function of galaxies, with a break at $r \approx R_c$, where R_c is the characteristic size of the ionized bubbles. Again using the excursion set formalism, we can estimate this modified bias in the limits $r \ll R_c$ and $r \gg R_c$.

By analogy with the halo model for the density field, these limiting regimes correspond to correlations between galaxies within a single bubble and within two separate bubbles. We begin with large scales: the observed clustering is the average bias of the bubbles weighted by the number of galaxies in each H II region (analogous to the two-halo term for the density field):

$$b_\infty = \int dm_b\, n_b(m_b)\, b_{\rm H\ II}(m_b)\, V_b \int_{m_D}^{\infty} dm_h\, \frac{n_h(m_h|m_b)}{\bar{n}_{\rm gal}}, \tag{11.43}$$

where we integrate only over those halos visible after damping wing absorption, $\bar{n}_{\rm gal}$ is the mean number density of observable galaxies, and $b_{\rm H\ II}(m_b)$ is the bias of an ionized bubble of mass m_b (see equation 9.24).

The behavior on small scales is somewhat more subtle. If galaxies were randomly distributed within each bubble, the simple argument in the first paragraph of this section suggests that the correlation function would just be the weighted average of the number of pairs per H II region. However, in addition to there being an increase in the number of galaxies in each bubble, the galaxies also trace density fluctuations within each bubble. On moderately small scales where nonlinear evolution in the density field may be neglected, we therefore write

$$b_{\rm sm}^2 = \int dm_b\, n_b(m_b)\, V_b\, b_h^2(m_b)\, \frac{\langle N_{\rm gal}(N_{\rm gal}-1)|m_b\rangle}{\bar{N}_{\rm gal}^2}, \tag{11.44}$$

where $\bar{N}_{\rm gal} = \bar{n}_{\rm gal} V_b$, $\langle N_{\rm gal}(N_{\rm gal}-1)|m_b\rangle$ is the expected number of galaxy pairs within each bubble, and b_h^2 measures the excess bias of these halos inside each bubble. Note the similarity to the halo-model calculation of the galaxy power spectrum here; in fact, this equation form can be derived formally by constructing the galaxy density field from bubbles and their constituent halos, in analogy to the halo model. This term then corresponds to the "two-halo, one-bubble" term in such a treatment; that is, correlations between two particles that lie in the same bubble but different dark matter halos. The "bubble profile" describing the distribution of galaxies within the bubble turns out to be proportional

to the square root of the linear matter correlation function. Provided the typical bubbles have more than two galaxies, we can write the expected number of pairs as

$$\langle N_{\rm gal}(N_{\rm gal}-1)|m_b\rangle \approx \max\{0,\, \bar{N}_{\rm gal}(m_b)[\bar{N}_{\rm gal}(m_b)-1]\}. \tag{11.45}$$

The remaining factor is $b_h(m_b)$. It may seem reasonable to take this to be the mean value of the usual excursion set halo bias, evaluated over $n(m_h|m_b)$. However, the pair density inside each bubble *already* includes much of this bias because it counts the number of galaxies in a region with overdensity $\delta_b = B$. We therefore want only the "excess" bias of the galaxies relative to density fluctuations on scales smaller than m_b, which is the bias evaluated from the conditional mass function in equation (11.37). Following the excursion set definition of this bias, we have

$$b_h(m_h|m_b) = 1 + \frac{(\delta_c-\delta_x)^2/(\sigma^2-\sigma_b^2)-1}{\delta_c(z=0)-\delta_x(z=0)}. \tag{11.46}$$

italics B(m subscript b) / AA (see eq. 11.37)

We show the resulting limiting bias cases at $z = 10$ as a function of $Q_{\rm H\,II}$ in Figure 11.18. In each panel, the different curves take different galaxy populations, with smaller galaxies having less net bias. Panels (a) and (b) show $b_{\rm sm}$ and b_∞. We scale the results to the bias $\bar{b}_h$ that would be intrinsic to the galaxy population if absorption could be ignored. Panel (c) shows the ratio $b_\infty/b_{\rm sm}$, illustrating the magnitude of the "break" in the linear bias. We emphasize that the scale at which the break occurs will evolve throughout reionization along with the characteristic bubble size R_c; for illustrative purposes we mark several values of R_c.

Clearly, both $b_{\rm sm}$ and b_∞ decrease throughout reionization. The large-scale bias decreases because the ionized regions must lie nearer the mean density (and hence be less biased) as $Q_{\rm H\,II} \to 1$: this behavior must be generic to any model in which reionization begins in overdense regions. The small-scale bias decreases because bubbles large enough to allow transmission become common: early on, only those galaxies with near neighbors are visible, so the correlations are strong. In the middle and final stages of reionization, most galaxies lie inside bubbles large enough to permit transmission, so more typical galaxies become visible, and $b_{\rm sm} \to \bar{b}_h$.

These qualitative results also hold true in more detailed calculations with numerical simulations. Figure 11.19 shows the estimated angular correlation function (i.e., the three-dimensional correlation function projected on the plane of the sky) from a radiative transfer simulation of LAEs at $z = 6.6$, the highest redshift window easily visible to a ground-based telescope. The different curves in each panel correspond to different ionized fractions; the different panels describe different surveys, with the top panel comparable to existing capabilities and the others a few times larger. Note the enhancement in small-scale correlations at small ionized fractions; this is the same effect we have described with $b_{\rm sm}$. The large-scale power is also enhanced, but it is much less sensitive to $Q_{\rm H\,II}$.

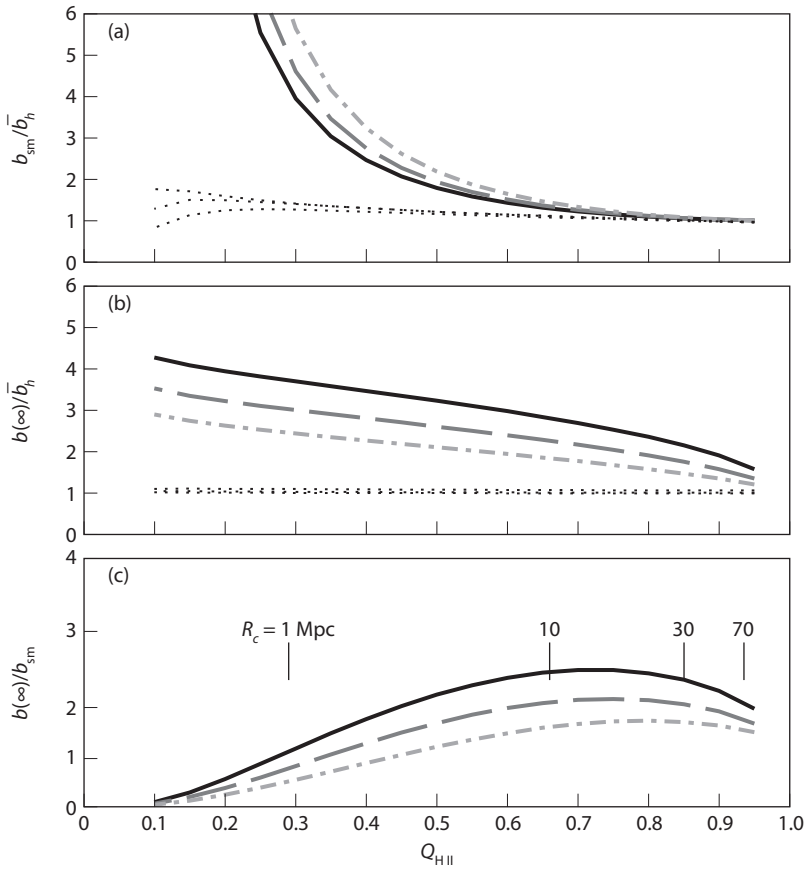


Figure 11.18 *(a)*: Predicted small-scale bias of LAEs at $z = 10$, relative to the bias that would be expected if all galaxies above the mass threshold were visible. This applies to separations larger than the nonlinear scale but smaller than the characteristic bubble size R_c. The solid, long-dashed, and short-dashed curves take $m_{\rm obs,min} = 10^8$, 10^9, and $10^{10}\,M_\odot$, respectively. The dotted curves show the predicted galaxy bias, neglecting absorption, relative to its true value (the small errors at early times result from the approximations described in the text). *(b)* Predicted large-scale bias ~~that would be~~ at $z = 10$, relative to the bias that would be expected if all galaxies above the mass threshold were visible. *(c)* Ratio of large-to-small scale bias; the transition between the two regimes occurs roughly at R_c, which is marked for a few different values of the bubble-filling factor $Q_{\rm H\,II}$. Furlanetto, S. R., et al., *Mon. Not. R. Astron. Soc.* **365**, 1012 (2006). Copyright 2006 by the Royal Astronomical Society.

PE

Although the correlation function and power spectrum (and through them the linear bias) are the most straightforward manifestations of the increased clustering, the "mask" applied to the galaxy distribution is itself non-Gaussian, so other clustering statistics—such as counts in cells or higher-order

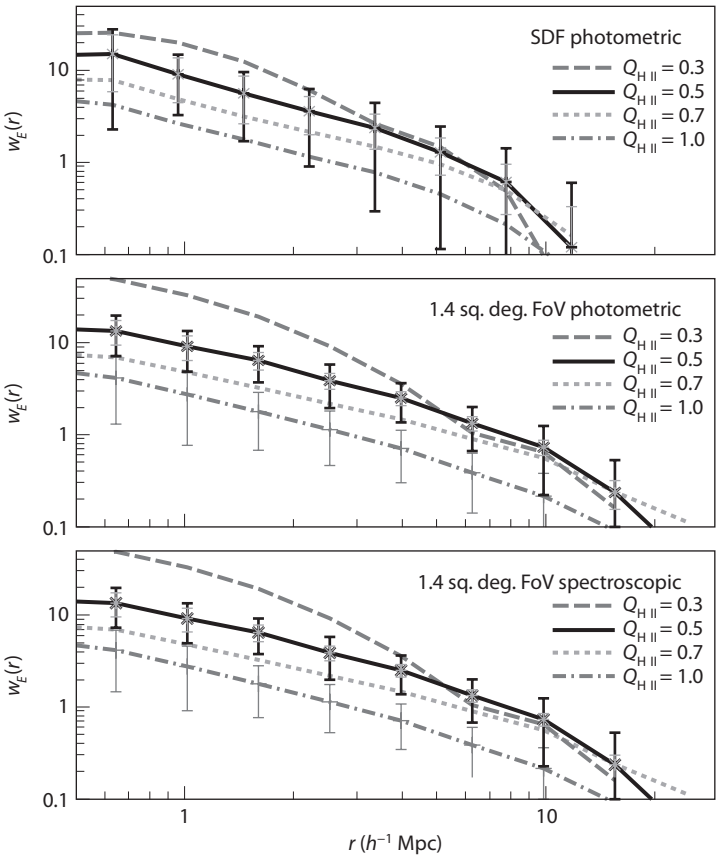


Figure 11.19 Angular correlation function of LAEs in a radiative transfer simulation of reionization. The simulation takes $z = 6.6$ and assumes all LAEs with an observed luminosity greater than the intrinsic luminosity of a halo with $m = 7 \times 10^{10}\,M\odot$ are visible. The different curves in each panel assume different ionized fractions. The top panel estimates the errors for existing surveys with the Subaru Deep Field in which LAEs are detected photometrically. The other two panels assume larger surveys (with $\sim$5 times more LAEs); the middle panel assumes a photometric survey, while the bottom one assumes the LAEs can be selected spectroscopically. In each one, the smaller error bars include Poisson fluctuations in the galaxy counts, while the larger spreads also include cosmic variance. McQuinn, M., et al., *Mon. Not. R. Astron. Soc.* **381**, 75 (2007). Copyright 2007 by the Royal Astronomical Society.

correlations—are also useful. All these probes follow the qualitative behavior of the bias, increasing most dramatically early in the reionization process.

Both the analytic and numerical approaches show that the bias increases rapidly with neutral fraction when $Q_{H\,II} < 0.5$, at least doubling and

Awkward spacing / PE

delete s / PE

sometimes increasing by an even larger amount, especially on large scales. This results, together with the change in the shape of the LAE correlation function with respect to the dark matter, makes the clustering signature much more robust to uncertainties in the nature of the LAE hosts. This is because the linear bias is a relatively slowly varying function of halo mass and redshift; mimicking the shift due to the ionized bubbles would require a drastic change in the properties of the galaxies.

However, it is worth emphasizing again that the radiative transfer of Lyman-α photons through the IGM is a complex process, and it can affect the observed clustering even after reionization is complete (thus the resonant absorption, which we have neglected in this section, can also be important). Interestingly, the frequency dependence of the scattering process induces anisotropies, generating clustering signatures analogous to redshift–space distortions. Fortunately, this component should not evolve as rapidly during reionization as the damping wing.[22]

Chapter Twelve

The 21-cm Line

As powerful as it is for studying the high-z Universe, the Lyman-α transition has the following major disadvantages:

- Most important, the Gunn-Peterson optical depth is enormous. Even a very small neutral fraction, of the order of $\sim 10^{-3}$, suffices to render the IGM opaque in this line. Thus, we are not able to use it to study the early, or even middle, phases of reionization except in special circumstances.
- Because the Lyman-α transition lies in the UV band, observing it requires bright UV sources that are very rare at high redshifts, limiting forest studies to only a modest number of lines of sight.
- The high excitation energy of the Lyman-α transition prevents us from using it to study the cold prereionization IGM, because the temperatures are much too low there to collisionally excite the line. Moreover, the large optical depth for absorption prevents us from measuring the IGM temperature through the line width.

The first of these difficulties can be overcome by using a resonant transition of a rarer element, such as a metal, but even if such elements, exist, their distribution introduces extra uncertainty into the interpretation (see §6.5). We can address all these problems by searching for a weaker, lower-energy line of atomic hydrogen: the best candidate is the *spin-flip* or hyperfine line. This transition was predicted theoretically by Hendrik van de Hulst in 1944 (following a suggestion by Jan Oort) and first observed from the sky by Harold Ewen and Ed Purcell through an office window at the Harvard Physics department in 1951. The transition is driven by the interaction of the spins of the proton and electron, whose relative directions affect the energy of the electron's orbit. An atom in the upper state eventually undergoes a spin-flip transition, emitting a photon with a wavelength of 21 cm. As we shall see, this transition is extremely weak, so the effective IGM optical depth is only of the order of 1%: this makes the entire neutral IGM accessible during the cosmic dawn. Moreover, the transition energy is so low that it provides a sensitive thermometer of the low-temperature IGM, and as a low-frequency radio transition, it can be seen across the entirety of the IGM against the CMB.

Figure 12.1 illustrates the power of the spin-flip transition with an analogy to the well-known structure of Swiss cheese. Each slice of cheese has a different structure, depending on where the air bubbles happen to lie within it. In the case of the spin-flip transition, by observing different wavelengths of $21(1 + z)$ cm, one slices the Universe at different redshifts z. Moreover,

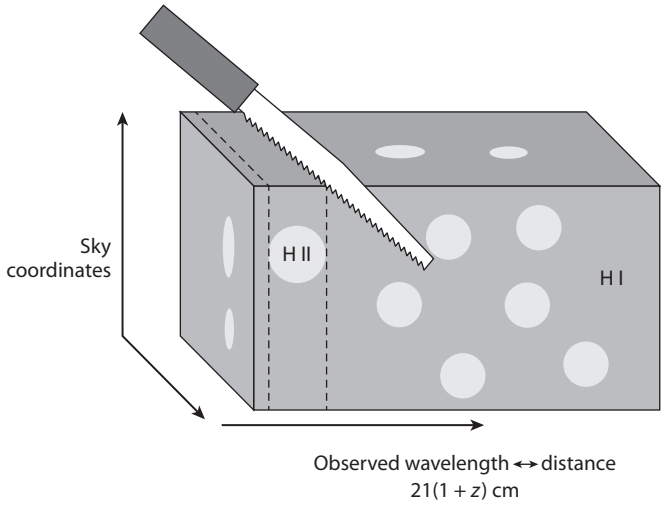


Figure 12.1 A diagram of 21-cm imaging of ionized bubbles during the epoch of reionization illustrating to analogy to slicing Swiss cheese. The technique of slicing at intervals separated by the typical dimension of a bubble is optimal for revealing different patterns in each slice.

the redshifted 21-cm emission should display angular structure as well as frequency structure owing to inhomogeneities in the gas density, the ionized fraction of hydrogen, and the fraction of excited atoms—the analogs of the air bubbles in Swiss cheese. A full map of the distribution of H I as a function of redshift would provide a three-dimensional image of the Swiss-cheese structure of the IGM during reionization. This mapping *tomography* provides the only way to map the distribution of >90% of the Universe's baryonic matter during the dark ages and cosmic dawn.

Figure 12.2, which shows a more concrete overview of the expected spin-flip signal, has two interesting aspects. The first is the sky-averaged, or monopole, brightness, which records the average properties of the H I as a function of observed wavelength (or, equivalently, cosmic time). This is shown in the bottom panel in brightness temperature units relative to the CMB (see later for a detailed discussion). Several different phases are labeled; we discuss each in turn in this chapter. The top panel shows the fluctuations inherent in this signal, which arise from the discrete, clustered luminous sources. The spin-flip background measures the UV and X-ray radiation fields over a broad swath of cosmic history, complementing the direct probes of individual galaxies that we have already described.

This chapter describes how we use the 21-cm line to study the high-z Universe. Following convention in the literature, we will often refer to the signal as "21-cm radiation," although in reality the *observed* wavelengths are larger by a factor of $(1 + z)$.

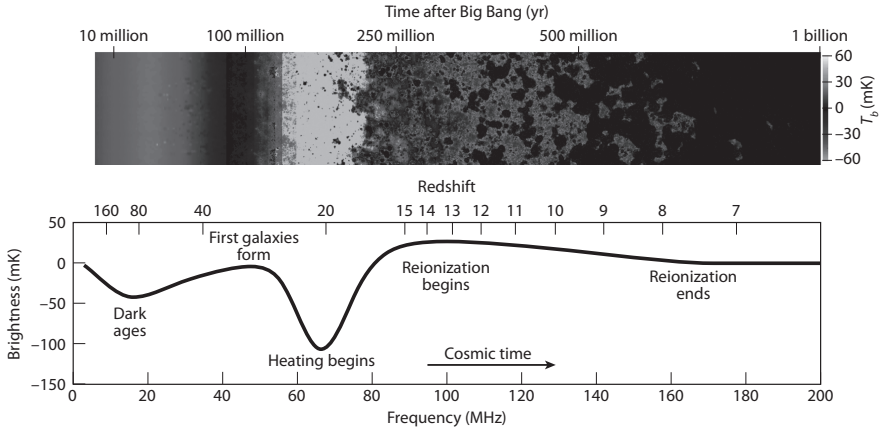


Figure 12.2 Overview of the expected 21-cm signal (see *Color Plate 26* for a color version of this figure). *Top panel:* Time evolution of fluctuations in the 21-cm brightness from just before the first stars form to the end of reionization. This evolution is pieced together from instantaneous redshift slices through a $(100\ \mathrm{Mpc})^3$ numerical simulation volume. Coloration indicates the strength of the 21-cm brightness as it transitions from absorption ~~(blue)~~ to emission ~~(red)~~ and finally disappears ~~(black)~~ owing to ionization. *Bottom panel:* Expected evolution of the sky-averaged 21-cm brightness from the dark ages at $z = 150$ to the end of reionization sometime before $z = 6$. The frequency structure is driven by the interplay of gas heating, the coupling of gas and 21-cm temperatures, and the ionization of the gas. The considerable uncertainty in the exact form of this signal arises from the poorly understood properties of the first galaxies. ~~Pritchard, J. R., & Loeb, A., *Nature* 468, 772 (2010). Copyright 2010 by Nature Publishing Group~~.

delete x3 / AA

12.1 Radiative Transfer of the 21-cm Line

The radiative transfer equation for the specific intensity I_ν of a spectral line is

$$\frac{dI_\nu}{d\ell} = \frac{\phi(\nu)h\nu}{4\pi}\left[n_1 A_{10} - (n_0 B_{01} - n_1 B_{10})\, I_\nu\right], \tag{12.1}$$

where $d\ell$ is a proper path length element, $\phi(\nu)$ is the line profile function normalized by $\int \phi(\nu)d\nu = 1$ (with an amplitude of the order of the inverse of the frequency width of the line and centered around the line frequency), subscripts 0 and 1 denote the lower and upper atomic levels, n_i denotes the number density of atoms at the different levels, and A_{ij} and B_{ij} are the Einstein coefficients for the transition between these levels (with i and j the initial and final states, respectively). In our case, the line frequency $\nu_{21} = 1420.4057$ MHz corresponds to a wavelength of $\lambda_{21} = 21.1061$ cm. We can then make use of the standard relations in atomic physics: $B_{10} = (g_0/g_1)B_{01}$, and $B_{10} = A_{10}(c^2/2h\nu^3)$, where g is the spin degeneracy factor of each state. For the 21-cm transition, $A_{10} = 2.85 \times 10^{-15}\ \mathrm{s}^{-1}$, and $g_1/g_0 = 3$.

Reprinted from *Nature*, 468, 772, Pritchard, J., & A. Loeb, "Cosmology: Hydrogen was not ionized abruptly," Fig. 1 (p. 772), Copyright 2010, with permission from Nature Publishing Group.

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

$$\left(\frac{n_1}{n_0}\right) = \left(\frac{g_1}{g_0}\right) \exp\left\{\frac{-T_*}{T_S}\right\}, \tag{12.2}$$

where $T_* \equiv E_{10}/k_B = 68\,\mathrm{mK}$ is equivalent to the transition energy E_{10}. In the regime of interest, $T_\star \ll T_\gamma$ as well as the spin temperature T_S, and so all related exponentials can be expanded to leading order.

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

In the Rayleigh-Jeans limit, the equation of radiative transfer along a line of sight through a cloud of uniform excitation temperature T_S becomes

$$T_b'(\nu) = T_S(1 - e^{-\tau_\nu}) + T_R'(\nu)e^{-\tau_\nu} \tag{12.3}$$

where the *optical depth* $\tau_\nu \equiv \int ds\,\alpha_\nu$ is the integral of the absorption coefficient (α_ν) along the ray through the cloud, T_R' is the brightness of the background radiation field incident on the cloud along the ray, and s is the proper distance. Because of the cosmological redshift, the emergent brightness $T_b'(\nu_0)$ measured in a cloud's comoving frame at redshift z creates an apparent brightness at Earth of $T_b(\nu) = T_b'(\nu_0)/(1+z)$, where the observed frequency is $\nu = \nu_0/(1+z)$. Henceforth, we will work in terms of these observed quantities.

The absorption coefficient is determined from the Einstein coefficients via[1]

$$\alpha = \phi(\nu)\frac{h\nu}{4\pi}(n_0 B_{01} - n_1 B_{10}). \tag{12.4}$$

Because all astrophysical applications have $T_S \gg T_*$, approximately three of four atoms find themselves in the excited state ($n_0 \approx n_1/3$). As a result, the stimulated emission correction is significant (and the net absorption depends on T_S).

In an expanding Universe with a local hydrogen number density $n_{\rm H}$ and with a velocity gradient along the line of sight of $dv_\parallel/dr_\parallel$, the 21-cm optical depth can be derived similarly to equation (4.11)[i]. Writing $\phi(\nu) \sim 1/(\Delta\nu)$ we obtain

$$\tau_{10} = \frac{3}{32\pi}\frac{hc^3A_{10}}{k_B T_S \nu_{10}^2}\frac{x_{\rm H\,I} n_{\rm H}}{(1+z)\,(dv_\parallel/dr_\parallel)} \tag{12.5}$$

$$\approx 0.0092\,(1+\delta)\,(1+z)^{3/2}\,\frac{x_{\rm H\,I}}{T_S}\left[\frac{H(z)/(1+z)}{dv_\parallel/dr_\parallel}\right], \tag{12.6}$$

In the second part T_S is Kelvin temperature, and we have scaled to the mean IGM density at z and to the average velocity gradient (the Hubble flow).

[i] Interestingly, the 21-cm case was computed by George Field in 1959, several years *before* the Gunn-Peterson calculation.[2]

In the latter case, $\Delta I_\nu \propto \Delta\ell\phi(\nu)\nu = |c\,dt/dz|(\nu\,dz/d\nu) = c/H$ and provides the analog of the Gunn-Peterson optical depth.

In practice, the background radiation source is usually the CMB, so $T_R' = T_\gamma(z)$, and we observe the contrast between high-redshift hydrogen clouds and the CMB. Because the optical depth is so small, we can expand the exponentials in equation (12.3) as follows:

$$T_b(\nu) \approx \frac{T_S - T_\gamma(z)}{1+z}\,\tau_{\nu_0} \tag{12.7}$$

$$\approx 9\,x_{\rm H\,I}(1+\delta)\,(1+z)^{1/2}\left[1-\frac{T_\gamma(z)}{T_S}\right]\left[\frac{H(z)/(1+z)}{dv_\parallel/dr_\parallel}\right]{\rm mK} \tag{12.8}$$

Here, $T_b < 0$ if $T_S < T_\gamma$, which yields an absorption signal. Emission occurs if $T_S > T_\gamma$; both regimes are important for the high-z Universe. Note that δT_b saturates if $T_S \gg T_\gamma$, but the absorption can become arbitrarily large if $T_S \ll T_\gamma$. The observability of the 21-cm transition therefore hinges on the spin temperature; we next describe the mechanisms that drive T_S either above or below $T_\gamma(z)$.

should be subscript lowercase greek gamma / PE

12.2 The Spin Temperature

Three competing processes determine T_S: *(i)* absorption of CMB photons (as well as emission stimulated by them); *(ii)* collisions with other hydrogen atoms, free electrons, and protons; and *(iii)* scattering of UV photons. In the presence of the CMB alone, the spin states reach thermal equilibrium with $T_S = T_\gamma$ on a timescale of $\sim T_*/(T_\gamma A_{10}) = 3\times10^5(1+z)^{-1}$ yr. This timescale is much shorter than the age of the Universe at all redshifts after cosmological recombination. However, the other two processes break this coupling. We let C_{10} and P_{10} be the de-excitation rates (per atom) from collisions and UV scattering, respectively. We also let C_{01} and P_{01} be the corresponding excitation rates. The spin temperature is then determined in equilibrium by[ii]

$$n_1\,(C_{10} + P_{10} + A_{10} + B_{10}I_{\rm CMB}) = n_0\,(C_{01} + P_{01} + B_{01}I_{\rm CMB})\,, \tag{12.9}$$

where $I_{\rm CMB}$ is the specific intensity of CMB photons. With the Rayleigh-Jeans approximation, we can rewrite equation (12.9) as

$$T_S^{-1} = \frac{T_\gamma^{-1} + x_c T_K^{-1} + x_\alpha T_c^{-1}}{1 + x_c + x_\alpha}, \tag{12.10}$$

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

$$\frac{C_{01}}{C_{10}} = \frac{g_1}{g_0}e^{-T_\star/T_K} \approx 3\left(1-\frac{T_\star}{T_K}\right). \tag{12.11}$$

[ii] Because the relevant timescales are all much shorter than the expansion time, equilibrium is an excellent approximation.

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

$$\frac{P_{01}}{P_{10}} \equiv 3\left(1 - \frac{T_\star}{T_c}\right). \tag{12.12}$$

We next calculate x_c, x_α, and T_c. In the limit in which $T_c \to T_K$ (a reasonable approximation in most situations of interest), we may write equation (12.10) as

$$1 - \frac{T_\gamma}{T_S} = \frac{x_c + x_\alpha}{1 + x_c + x_\alpha}\left(1 - \frac{T_\gamma}{T_K}\right). \tag{12.13}$$

12.2.1 Collisional Coupling

We first consider collisional excitation and de-excitation of the hyperfine levels, which dominate in dense gas. The coupling coefficient for collisions with species i is

$$x_c^i \equiv \frac{C_{10}^i}{A_{10}}\frac{T_\star}{T_\gamma} = \frac{n_i\,\kappa_{10}^i}{A_{10}}\frac{T_\star}{T_\gamma}, \tag{12.14}$$

where κ_{10}^i is the rate coefficient for spin de-excitation in collisions with that species (with units of $\mathrm{cm^3\,s^{-1}}$). The total x_c is the sum over all species i, which in principle includes collisions with (*i*) other hydrogen atoms, (*ii*) free electrons, (*iii*) protons, and (*iv*) other species (helium and deuterium); the last turn out to be unimportant.

These rate coefficients are ultimately determined by the quantum-mechanical cross sections of the relevant processes. We do not list them in detail but merely present the results in Figure 12.3.[3] Although the atomic cross section is small, in the unperturbed IGM, collisions between neutral hydrogen atoms nearly always dominate these rates, because the ionized fraction is small. Free electrons can be important in partially ionized gas; collisions with protons are important only at the lowest temperatures.

Crucially, the collisional coupling is quite weak in a nearly neutral, cold medium. Thus, the overall density must be large for this process to effectively fix T_S. A convenient estimate of their importance is the critical overdensity, $\delta_{\rm coll}$, at which $x_c = 1$ for H–H collisions:

$$1 + \delta_{\rm coll} = 0.99\left[\frac{\kappa_{10}(88\ \mathrm{K})}{\kappa_{10}(T_K)}\right]\left(\frac{0.023}{\Omega_b h^2}\right)\left(\frac{70}{1+z}\right)^2, \tag{12.15}$$

where we have inserted the expected temperature at $1 + z = 70$. In the standard picture, at redshifts $z < 70$, $x_c \ll 1$, and $T_S \to T_\gamma$; by $z \sim 30$ the IGM essentially becomes invisible. It is worth emphasizing, however, that κ_{10} is extremely sensitive to T_K in this regime. If the universe is somehow heated above the fiducial value, the threshold density can remain modest: $\delta_{\rm coll} \approx 1$ at $z = 40$ if $T_K = 300\ \mathrm{K}$.

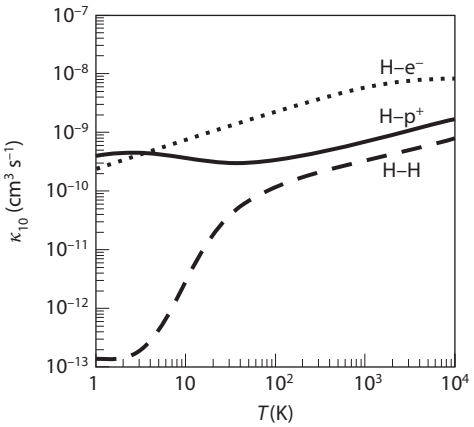


Figure 12.3 De-excitation rate coefficients for H–H collisions (dashed line), H–e$^-$ collisions (dotted line), and H–p$^+$ collisions (solid line). Note that the net rates are also proportional to the densities of the individual species, so H–H collisions still dominate in a weakly ionized medium. Furlanetto, S. R., & Furlanetto, M. R., *Mon. Not. R. Astron. Soc.* **379**, 130 (2007). Copyright 2007 by the Royal Astronomical Society.

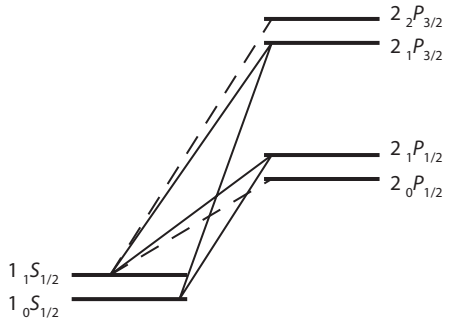


Figure 12.4 Level diagram illustrating the Wouthuysen-Field effect. We show the hyperfine splittings of the $1S$ and $2P$ levels. The solid lines label transitions that mix the ground-state hyperfine levels, while the dashed lines label complementary allowed transitions that do not participate in mixing. Pritchard, J. R., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006). Copyright 2006 by the Royal Astronomical Society.

12.2.2 The Wouthuysen-Field Effect

We therefore require a different mechanism to break the coupling to the CMB during the era of the first galaxies, namely, the *Wouthuysen-Field mechanism* (named after the Dutch physicist Siegfried Wouthuysen and Harvard astrophysicist George Field who first explored it).[4] The mechanism is illustrated in Figure 12.4, where we have drawn the hyperfine sublevels of the $1S$ and $2P$

states of H I. Suppose a hydrogen atom in the hyperfine singlet state absorbs a Lyman-α photon. The electric dipole selection rules allow $\Delta F = 0, 1$ except that $F = 0 \rightarrow 0$ is prohibited (here F is the total angular momentum of the atom). Thus the atom will jump to either of the central $2P$ states. However, these same rules allow this state to decay to the ${}_1S_{1/2}$ triplet level.[iii] Thus, atoms can change hyperfine states through the absorption and spontaneous reemission of a Lyman-α photon (or, indeed, any Lyman-series photon).

The Wouthuysen-Field coupling must depend on the total rate (per atom) at which Lyman-α photons are scattered within the gas,

$$P_\alpha = 4\pi\sigma_0 \int d\nu\, J_\nu(\nu)\phi_\alpha(\nu), \tag{12.16}$$

where $\sigma_\nu \equiv \sigma_0\phi_\alpha(\nu)$ is the local absorption cross section, $\sigma_0 \equiv (\pi\, e^2/m_e\, c)\, f_\alpha$, $f_\alpha = 0.4162$ is the oscillator strength of the Lyman-α transition, $\phi_\alpha(\nu)$ is the Lyman-α absorption profile, and J_ν is the angle-averaged specific intensity of the background radiation field.[iv] The line typically has a Voigt profile ϕ_V, as described in §11.1.1.

What about transitions to higher Lyman-n levels? Suppose that a photon redshifts into the Lyman-n resonance. After absorption, it can either scatter (through a decay directly to the ground state) or cascade through a series of intermediate levels and produce different photons. The direct decay probabilities are $P_{nP\rightarrow 1S} \sim 0.8$, so a Lyman-$n$ photon will typically scatter $N_{\rm scatt} \approx (1 - P_{nP\rightarrow 1S})^{-1} \sim 5$ times before instead initiating a decay cascade. In contrast, Lyman-α photons scatter hundreds of thousands of times before being destroyed (see §11.1.1 for some examples). As a result, coupling from the direct scattering of Lyman-n photons is suppressed compared with Lyman-α by a large factor.

However, Lyman-n photons can still be important because of their cascade products. Consider the decay chains shown in Figure 12.5. For Lyman-β, the only permitted decays are to the ground state (which regenerates a Lyman-β photon and starts the process again) or to the $2S$ level. The Hα photon produced in the $3P \rightarrow 2S$ transition (and, indeed, any photon produced in a decay to an excited state) escapes to infinity. Thus, the atom eventually finds itself in the $2S$ state, which decays to the ground state via a forbidden two-photon process with $A_{2S\rightarrow 1S} = 8.2\,{\rm s}^{-1}$. These photons, too, escape to infinity. Thus coupling from Lyman-β photons can be completely neglected.[v]

But, now, consider Lyman-γ excitation, also shown in Figure 12.5. This can cascade (through $3S$ or $3D$) to the $2P$ level, in which case the original Lyman-n photon is "recycled" into a Lyman-α photon, which then scatters many times through the IGM. Thus, the key quantity for determining the coupling induced

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

[iv] By convention, we use the specific intensity in units of photons cm^{-2} Hz^{-1} s^{-1} sr^{-1} here, which is better conserved during the expansion of the Universe (whereas energy redshifts away).

[v] In a medium with very high density, atomic collisions can mix the two $n = 2$ angular momentum states, but that process is unimportant in the IGM.

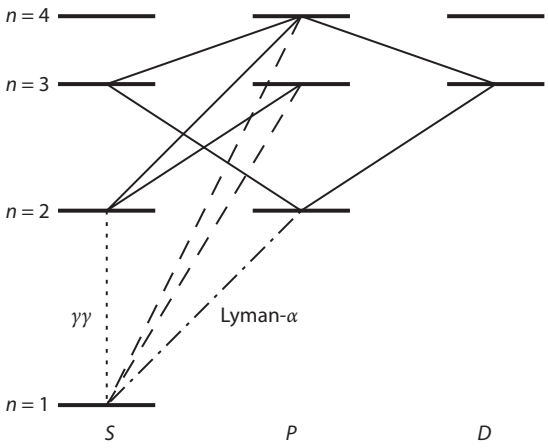


Figure 12.5 Decay chains for Lyman-β and Lyman-γ excitations. We show Lyman-n transitions by dashed curves, Lyman-α by the dot-dashed curve, cascades by solid curves, and the forbidden $2S \rightarrow 1S$ transition by the dotted curve. Pritchard, J. R., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006). Copyright 2006 by the Royal Astronomical Society.

by Lyman-n photons is the fraction $f_{\rm rec}(n)$ of cascades that terminate in Lyman-α photons. We have seen that $f_{\rm rec}(n = 3)$ vanishes, but detailed quantum-mechanical calculations show that the higher states all have $f_{\rm rec} \sim 1/3$.[5]

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

$$P_{01} \propto B_{\rm ad} J_{\rm ad} \frac{A_{\rm db}}{A_{\rm da} + A_{\rm db}} + B_{\rm ae} J_{\rm ae} \frac{A_{\rm eb}}{A_{\rm ea} + A_{\rm eb}}. \tag{12.17}$$

The first term contains the probability for an a$\rightarrow$d transition ($B_{\rm ad} J_{\rm ad}$), together with the probability that the subsequent decay will terminate in state b; the second term is the same for transitions to and from state e. Next, we need to relate the individual A_{ij} coefficients to $A_\alpha = 6.25 \times 10^8$ Hz, the total Lyman-α spontaneous emission rate (averaged over all the hyperfine sublevels). We can do this by using a sum rule stating that the sum of decay intensities ($g_i A_{ij}$) for transitions from a given nFJ to all the $n'J'$ levels (summed over F') is proportional to $2F + 1$;[6] the relative strengths of the permitted transitions are then (1, 1, 2, 2, 1, 5), where we have ordered the lines (bc, ad, bd, ae, be, bf), and the two-letter labels represent the initial and final states. With our assumption that the background radiation field is constant across the individual hyperfine lines, we find $P_{10} = (4/27) P_\alpha$.

We may then write as the coupling coefficient x_α as

$$x_\alpha = \frac{4P_\alpha}{27A_{10}}\frac{T_\star}{T_\gamma} = S_\alpha \frac{J_\alpha}{J_\nu^c}, \tag{12.18}$$

where in the second equality we evaluate J_ν at line center and set $J_\nu^c \equiv 1.165 \times 10^{-10}[(1+z)/20]\,\mathrm{cm^2\,s^{-1}\,Hz^{-1}\,sr^{-1}}$. We include here a correction factor S_α that accounts for variations in the intensity near the line center (see later discussion). This coupling threshold for $x_\alpha = S_\alpha$ can also be written in terms of the number of Lyman-α photons per hydrogen atom in the Universe, which we denote $\tilde{J}_\nu^c = 0.0767\,[(1+z)/20]^{-2}$. This threshold is relatively easy to achieve in practice.

Two challenges remain: to calculate T_c and the correction factor S_α. The former is the effective temperature of the UV radiation field, defined in equation (12.12), which is determined by the shape of the photon spectrum at the Lyman-α resonance. That the effective temperature of the radiation field *must* matter is easy to see: the energy deficit between the different hyperfine splittings of the Lyman-α transition (labeled bc, ad, etc.) implies that the mixing process is sensitive to the gradient of the background spectrum near the Lyman-α resonance. More precisely, the procedure described after equation (12.17) yields

$$\frac{P_{01}}{P_{10}} = \frac{g_1}{g_0}\frac{n_{\rm ad}+n_{\rm ae}}{n_{\rm bd}+n_{\rm be}} \approx 3\left(1 + v_0\frac{d\ln n_\nu}{d\nu}\right), \tag{12.19}$$

where $n_\nu = c^2 J_\nu/2\nu^2$ is the photon occupation number. Thus, by comparison with equation (12.12) we find

$$\frac{h}{k_B T_c} = -\frac{d\ln n_\nu}{d\nu}. \tag{12.20}$$

A simple argument shows that $T_c \approx T_K$:[7] so long as the medium is extremely optically thick, the enormous number of Lyman-α scatterings must bring the Lyman-α profile to a blackbody of temperature T_K near the line center. This condition is easily fulfilled in the high-redshift IGM, where $\tau_\alpha \gg 1$. Specifically, atomic recoils during scattering (the last term in equation 11.9) tilt the spectrum to the red and are primarily responsible for establishing this equilibrium.

The scattering process is actually much more complicated than naively expected, because scattering itself modifies the shape of J_ν. Intuitively, a flat input spectrum develops an absorption feature because of the increased scattering rate near the Lyman-α resonance. Photons continually lose energy by redshifting, but they also lose energy through recoil whenever they scatter. If the fractional frequency drift rate is denoted by $\mathcal{A}$, continuity requires that $n_\nu\mathcal{A}$ = constant; when $\mathcal{A}$ increases near resonance, the number density must fall. On average, the energy loss (or gain) per scattering is[8]

$$\frac{\Delta E_{\rm recoil}}{E} = \frac{h\nu}{m_p c^2}\left(1 - \frac{T_K}{T_c}\right), \tag{12.21}$$

where the first factor comes from recoil off an isolated atom, and the second factor corrects for the distribution of initial photon energies; the energy loss

vanishes when $T_c = T_K$, and when $T_c < T_K$, the gas is heated by the scattering process.

To compute the suppression factor in the intensity we must calculate the photon spectrum near Lyman-α. We begin with the radiative transfer equation in an expanding universe (written in comoving coordinates and again using units of $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{Hz}^{-1}\,\mathrm{sr}^{-1}$ for J_ν; cf. equation 4.42):

$$\frac{1}{cn_{\rm H}\sigma_0}\frac{\partial J_\nu}{\partial t} = -\phi_\alpha(\nu)\,J_\nu + H\nu_\alpha\frac{\partial J_\nu}{\partial \nu} + \int d\nu'\,R(\nu,\nu')\,J_{\nu'} + C(t)\psi(\nu). \quad (12.22)$$

Here the first term on the right-hand side describes absorption, the second is redshifting due to the Hubble flow, and the third accounts for reemission following absorption. $R(\nu,\nu')$ is the "redistribution function" that describes the frequency of an emitted photon, which depends on the relative momenta of the absorbed and emitted photons as well as the absorbing atom. The last term describes injection of new photons: C is the rate at which they are produced, and $\psi(\nu)$ is their frequency distribution.

The redistribution function R is the complicated aspect of the problem, but it can be simplified if the frequency change per scattering (typically of the order of $\Delta\nu_D$) is "small." In that case, we can expand $J_{\nu'}$ to second order in $(\nu-\nu')$ and rewrite equation (12.22) as a diffusion problem in frequency. The steady-state version of equation (12.22) becomes, in this so-called Fokker-Planck approximation,

$$\frac{d}{dx}\left(-\mathcal{A}\,J + \mathcal{D}\frac{dJ}{dx}\right) + C\psi(x) = 0, \quad (12.23)$$

where $x \equiv (\nu-\nu_\alpha)/\Delta\nu_D$, $\mathcal{A}$ is the frequency drift rate, and $\mathcal{D}$ is the diffusivity. In general, the Fokker-Planck approximation is valid when (*i*) the frequency change per scattering ($\sim\Delta\nu_D$) is smaller than the width of any spectral features, and either (*iia*) the photons are outside the line core, where $d\phi_\alpha/dx$ is small, or (*iib*) the atoms are in equilibrium with $T_c \approx T_K$.

Solving for the background spectrum thus reduces to specifying $\mathcal{A}$ and $\mathcal{D}$. The first involves the Hubble flow, which causes a drift $\mathcal{A}_H = -\tau_\alpha^{-1}$ (without any associated diffusion). The remaining terms come from R and incorporate all the physical processes relevant to energy exchange in scattering. The drift from recoil is[9]

$$\mathcal{D}_{\rm scatt} = \phi_\alpha(x)/2, \quad (12.24)$$

$$\mathcal{A}_{\rm scatt} = -(\eta - x_0^{-1})\phi_\alpha(x), \quad (12.25)$$

where $x_0 \equiv \nu_\alpha/\Delta\nu_D$, and $\eta \equiv (h\nu_\alpha^2)/(m_p c^2\Delta\nu_D)$. The latter is the recoil parameter measuring the average loss per scattering in units of the Doppler width.

Finally, to solve equation (12.23) we must specify the boundary conditions, which essentially correspond to the input photon spectrum (ignoring scattering) and the source function. Because the frequency range of interest is so narrow, two cases suffice: a flat input spectrum (which approximately describes photons that redshift through the Lyman-α resonance, regardless of the initial

source spectrum) and a step function, where photons are "injected" at line center (through cascades or recombinations) and redshift away. In either case, the first integral over x is trivial. At high temperatures, where spin flips are unimportant to the overall energy exchange, we can write

$$\phi \frac{dJ}{dx} + 2\{[\eta - (x + x_0)^{-1}]\phi + \tau_\alpha^{-1}\}J = 2K/\tau_\alpha. \quad (12.26)$$

The integration constant K equals J_∞, the flux far from resonance, both for photons that redshift into the line and for injected photons at $x < 0$; it is zero for injected photons at $x > 0$.

The formal analytic solution, when $K \neq 0$, is most compactly written in terms of $\delta_J \equiv (J_\infty - J)/J_\infty$:[vi]

$$\delta_J(x) = 2\eta \int_0^\infty dy \exp\left[-2\{\eta - (x + x_0)^{-1}\}y - \frac{2}{\tau_\alpha}\int_{x-y}^{x} \frac{dx'}{\phi_\alpha(x')}\right]. \quad (12.27)$$

(An analogous form also exists for photons injected at line center.) The full problem, including the intrinsic Voigt profile of the Lyman-α line, must be solved numerically,[10] but including only the Lorentzian wings from natural broadening allows a simpler solution that is quite accurate in the most interesting regime of $T_K < 1{,}000$ K.

The crucial aspect of equation (12.27) is that (as expected from the preceding qualitative argument) an absorption feature appears near the line center; its strength is roughly proportional to η, our recoil parameter. The feature is more significant when T_K is small (or the average effect of recoil is large). Figure 12.6 shows some example spectra (both for a continuous background and for photons injected at line center).

Usually, the most important result is the suppression of the radiation spectrum at line center compared with the assumed initial condition. This decreases the total scattering rate of Lyman-α photons (and hence the Wouthuysen-Field coupling) below what one naively expects. The suppression factor (from equation 12.18) is

$$S_\alpha = \int_{-\infty}^{\infty} dx\, \phi_\alpha(x)\, J(x) \approx [1 - \delta_J(0)] \leq 1, \quad (12.28)$$

where the second equality follows from the narrowness of the line profile. Again, the Lorentzian wing approximation turns out to be an excellent one; when $T_K \gg T_\star$, the suppression is

$$S_\alpha \sim \exp\left[-0.803\left(\frac{T_K}{1\text{ K}}\right)^{-2/3}\left(\frac{\tau_\alpha}{10^6}\right)^{1/3}\right]. \quad (12.29)$$

Note that this form applies to both photons injected at line center as well as those that redshift in from infinity. As we can see in Figure 12.6, the suppression is most significant in cool gas.

[vi] Here we assume the gas has a sufficiently high temperature that the different hyperfine subtransitions can be treated as one.

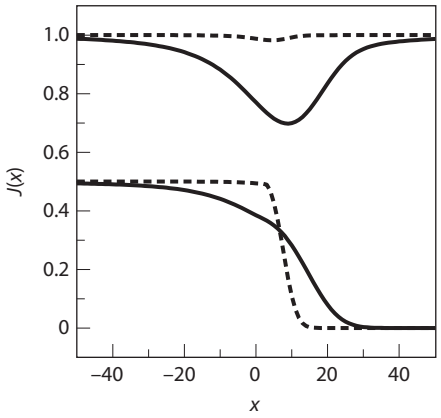


Figure 12.6 Background radiation field near the Lyman-α resonance at $z = 10$; $x \equiv (\nu - \nu_\alpha)/\Delta\nu_D$ is the normalized deviation from line center. The upper and lower sets are for continuous photons and photons injected at line center, respectively. (The former are normalized to J_∞; the latter have arbitrary normalization.) The solid and dashed curves take $T_K = 10$ and 1,000 K, respectively. Furlanetto, S. R., & Pritchard, J. R., *Mon. Not. R. Astron. Soc.* **372**, 1093 (2006). Copyright 2006 by the Royal Astronomical Society.

The fundamental reason for the suppression is the recoil from scattering. Momentum conservation during each scattering slightly decreases the frequency of the photon. The strongly enhanced scattering rate near line center means that photons "flow" through that region of the spectrum more rapidly than elsewhere (where only the cosmological redshift applies), so the amplitude of the spectrum is smaller. Meanwhile, the scattering in such an optically thick medium also causes photons to diffuse away from line center, broadening the feature well beyond the nominal line width.

12.3 The Brightness Temperature of the Spin-Flip Background

With the basic atomic physics of the 21-cm line in place, we now turn to estimating the astrophysical inputs that determine its properties. Of course, these inputs are at the moment unknown, so at first we keep the discussion general and then later focus on some particular simple models as examples.

12.3.1 Feedback: The Lyman-α Background

After $z \sim 30$, when collisional coupling becomes unimportant, the spin temperature is determined by the scattering of Lyman-α photons. In practice, the relevant photons do not start at the Lyman-α wavelength, because those redshift out of resonance very soon after they are created and do not contribute to the

coupling except very near their sources. Instead, the important photons begin in the UV and redshift into a Lyman-series line, possibly cascading down to a Lyman-α photon.

To compute J_α, we therefore begin with the proper UV emissivity at a frequency ν, $\epsilon(\nu, z)$. Here we consider the simple limit in which this emissivity is nearly uniform.

In fact, we have already discussed this background in some detail, for these photons, which range in energy from 10.2 to 13.6 eV, are (nearly) the same as those that contribute to the Lyman-Werner background that dissociated H_2 molecules in the early Universe (see §6.1.4). The difference is that here we are concerned not with the photons between the Lyman resonances but with those photons that do redshift into those resonances (and then cascade into Lyman-α, in the case of higher-n transitions). Given the UV emissivity, the desired background is

$$\begin{aligned} J_\alpha(z) &= \sum_{n=2}^{n_{\rm max}} J_\alpha^{(n)}(z) \\ &= \frac{c}{4\pi} \sum_{n=2}^{n_{\rm max}} f_{\rm rec}(n) \int_z^{z_{\rm max}(n)} dz' \left|\frac{dt}{dz'}\right| \left(\frac{1+z}{1+z'}\right)^3 4\pi \frac{c}{H(z')} \epsilon(\nu_n', z'), \end{aligned} \tag{12.30}$$

where ν_n' is the frequency at redshift z' that redshifts into the Lyman-n resonance at redshift z, and $z_{\rm max}(n)$ is the largest redshift from which a photon can redshift into the Lyman-n resonance. This equation is very similar to equation (6.7)—and hence equation (4.43)—because the Lyman-Werner and Wouthuysen-Field backgrounds arise from nearly the same photons. The only major difference is that the latter includes photons that cascade from higher Lyman-series absorptions, so it requires a sum over all those line frequencies. The sum must be truncated at some large $n_{\rm max}$ that is determined by the typical size of ionized regions around the sources, but the result is not sensitive to the precise cutoff value.

Just as with the Lyman-Werner background, the Lyman-α intensity is fairly uniform in the standard cosmological model: in fact the effective "horizon" within which a given source is visible is even larger than in that other case, because the gap between Lyman-α and Lyman-β corresponds to $\sim$250 comoving Mpc. However, unlike for the Lyman-Werner background, Wouthuysen-Field coupling is rather sensitive to the precise intensity of the background, so the fluctuations are still very important. Moreover, just as for the Lyman-Werner background, this horizon is comparable to the scales over which the relative baryon and dark matter velocities vary, so that may induce much stronger fluctuations in the Wouthuysen-Field coupling (see §3.2.2).[11] For simplicity, we ignore the latter effect here.

These fluctuations, in turn, depend on the sources of the photons, most likely star-forming galaxies. If the star formation rate traces the rate at which matter

collapses into galaxies, the comoving emissivity at frequency ν is

$$\epsilon_c(\nu, z) = f_\star \frac{\rho_b}{m_p} N_{\mathrm{L}n}(\nu) \frac{df_{\mathrm{coll}}}{dt}, \tag{12.31}$$

where $N_{\mathrm{L}n}(\nu)$ is the number of photons produced in the frequency interval $\nu \pm d\nu/2$ per baryon incorporated into stars. Although real spectra are rather complicated, a useful quantity is the total number N_α of photons per baryon in the interval 10.2–13.6 eV (which is very similar to N_{LW} in equation 6.10). For low-metallicity Population II stars and very massive Population III stars, these values are $N_\alpha = 9{,}690$ and $N_\alpha = 4{,}800$, respectively.[12]

Of course, processes other than star formation can also create a Lyman-α background. These include especially UV photons from quasars (which can be modeled in the same way at stars, though $N_{\mathrm{L}n}$ changes) and collisional excitation by higher-energy X-rays. For the latter, a fraction $f_c \sim f_i \sim x_{\mathrm{H\,I}}/3$ of the energy is typically lost to excitations, and ≈ 0.8 of that energy ends up in Lyman-α photons.[13] It is easy to see that it can be quite significant. The critical intensity for the Wouthuysen-Field effect corresponds to $\sim$1 photon per 10 hydrogen atoms, or $\sim$1 eV per atom. Because the fractions of energy deposited as ionization and collisional excitation are comparable, any scenario that appeals to X-rays for significant ionization would also induce strong coupling once the IGM became $\sim$10% ionized.

In any case, computing the fluctuations in the intensity is a more difficult task. The simplest approach is to use a modified version of the halo model (introduced in §3.6.1) applied to the radiation background instead of to the density field. Here, we construct the background by imagining that each galaxy is surrounded by a radiation field with a specified shape, $J_h(r|\Psi)$, where Ψ labels all the parameters that may determine an individual galaxy's luminosity (principally, we presume later, the host halo's mass). The total radiation background $J(\mathbf{x})$ is then the sum of the radiation from all the halos, just as the density field is the sum of the density profiles of each dark matter clump in the halo model. We can therefore use the usual machinery of the halo model to describe the radiation background. Then,

$$P_J(k) = P_J^{1h}(k) + P_J^{2h}(k), \tag{12.32}$$

where P_J^{1h} describes correlations from a single galaxy's radiation, and P_J^{2h}, those between galaxies.

The key input is therefore to determine the intensity profile of each galaxy. If we consider only the radiation between Lyman-α and Lyman-β, the profile of photons that redshift into the Lyman-α resonance will follow the usual $1/r^2$ law, with just two modifications: (*i*) the profile will be truncated where those photons with the largest initial energies (just below Lyman-β) redshift into the Lyman-α resonance, and (*ii*) the relevant emitted frequency (chosen so that it redshifts into Lyman-α resonance at r) varies with radius r. However, we must also add those photons that redshift into a higher Lyman-n series resonance

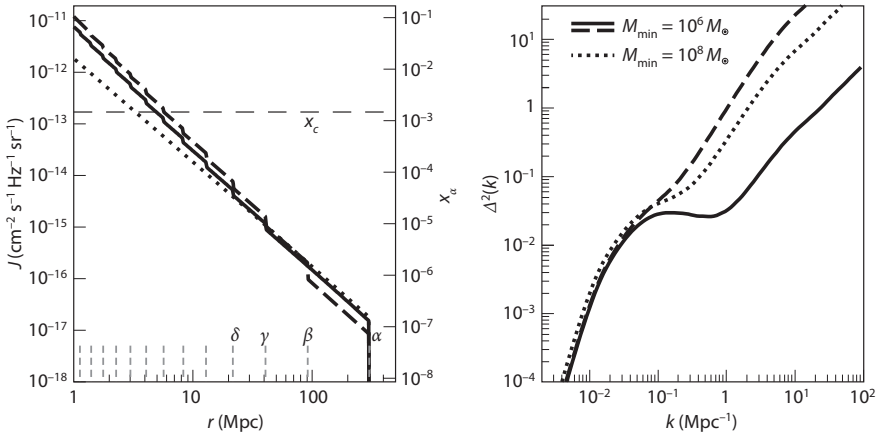


Figure 12.7 *Left:* Lyman-α flux profile of a massive galaxy at $z = 20$. The solid line shows the full calculation, the dashed line assumes $f_{\rm rec} = 1$ for all n, and the dotted line ignores all cascades from higher energies (all are normalized to have the same total luminosity). The vertical dashed lines along the horizontal axis mark the horizons of the respective Lyman transitions. The right axis converts the local flux to the Wouthuysen-Field coupling coefficient assuming $S_\alpha = 1$; for context, the dashed horizontal line shows the collisional coupling coefficient at $z = 20$ assuming no IGM heating. *Right:* Dimensionless power spectrum of the Lyman-α background for several simple star formation scenarios. The solid and dashed curves are chosen to be near the peak of the Lyman-α fluctuations in each scenario. They assume $M_{\rm min} = 10^6$ $M_\odot$ at $z = 30$ (~~bottom~~ curve) and $M_{\rm min} = 10^8$ $M_\odot$ at $z = 20.5$ (~~top~~ curve); both take $f_\star = 0.1$ to normalize the background. The dashed curve takes $M_{\rm min} = 10^6$ $M_\odot$ at $z = 30$ but assumes that each halo can form stars for only 3 Myr, ~2% of the age of the Universe. Pritchard, J. R., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006); copyright 2006 by the Royal Astronomical Society. Holzbauer, L. N., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.*, **419**, 718 (2012). Copyright 2011 by the Royal Astronomical Society.

solid / dotted / AA

and then cascade to Lyman-α. Thus, the total profile is

$$J_{h,\alpha}(r) = \sum_{n=1}^{\infty} f_{\rm rec}(n) \frac{L(\nu'_n|\Psi)/h\nu'_n}{(4\pi r)^2}, \tag{12.33}$$

where $L(\nu|\Psi)$ is the luminosity per unit frequency from the source (with parameters Ψ), ~~and~~ ν'_n is the frequency that redshifts into the Lyman-n resonance at r, and each term in the sum is included only when r is smaller than the effective horizon for these photons. The left panel in Figure 12.7 shows this profile for a massive galaxy at $z = 20$. Note that it is slightly steeper than the $1/r^2$ expectation at moderate distances from the source, owing to the cascade effects.

PE

In practice, the light travel time over the Lyman-α horizon is >100 Myr, a substantial fraction of both the age of Universe and the lifetime of a typical source (either stars or quasars). As a result, the source luminosity likely changes significantly during the time period of interest, and some estimate of source evolution is necessary. In the simplest models, one can take $L \propto df_{\rm coll}/dt$ in order to reflect the overall evolution of gas inside galaxies.

The right panel of Figure 12.7 shows some example power spectra for the intensity of this background. We consider two cases: one in which the luminous sources are extremely small (existing in all halos above $10^6\ M_\odot$, solid curves) and one in which only massive halos host sources (above $10^8\ M_\odot$, ~~dashed~~ dotted curves). In both cases, we have normalized the mean amplitude of the Lyman-α background so that the background just reaches the coupling threshold, $x_\alpha \approx 1$,[vii] by setting $f_\star = 0.1$ and $z = 20$ and $z = 30$ for the high- and low-mass case, respectively. In both cases, intensity fluctuations are small on large scales: <1% for $k < 0.1\,{\rm Mpc}^{-1}$, comparable to the horizon of each source.

If sources are common, the fluctuations remain small at smaller scales as well, owing to the enormous number of them. But if sources are rare, the "one-halo" term representing the intensity profile of each source becomes important at moderate or large scales, and the fluctuations can be moderately large on these scales. The dashed curve illustrates this situation: it assumes that each halo can form stars only in a single burst lasting 3 Myr (such as if each halo hosts a single short burst of Population III star formation). In this case, the fluctuations are much stronger, because at any given time most of the halos are invisible (in the language of the halo model, the occupation fraction is very small even above the minimum mass for star formation), but the average radiation background also falls well below threshold. As mentioned previously, the fluctuations also increase substantially if the relative streaming of baryons and dark matter is included.

In §12.5 we consider how fluctuations in this background translate into fluctuations in the 21-cm signal.

12.3.2 Feedback: IGM Heating

The Wouthuysen-Field background couples the spin temperature to the gas kinetic temperature, so we must also compute the latter. A number of processes may contribute to it: shock heating from structure formation, UV photons, and X-rays.

The role of shock heating is unclear: very little of the IGM gas had been incorporated into sheets or filaments at these times, so the usual shocks that surround the cosmic web are unimportant. However, the very low temperature gas ($T < 30$ K) has large peculiar velocities from gravitational infall, and shocks may have occured earlier in such an environment. The large-scale baryon velocities generated during recombination (see §2.1.2) may be important in this regard.

[vii] More precisely, these curves have $\beta_\alpha = 1$ (see equation 12.45).

dotted / AA

insert space after hyphen / PE

r / PE

However, we have seen that ionizing photons very effectively heat the gas, but the ionized gas itself does not, of course, contribute to the 21-cm signal.[viii]

The photons that trigger Lyman-α coupling do exchange energy with the IGM, through recoil. The typical energy exchange per scattering is small (see equation 12.21), but the number of scatterings is large. If the net heating rate per atom followed the naive expectation, $\sim P_\alpha \times (h\nu_\alpha)^2/m_p c^2$, the gas temperature would exceed T_γ soon after Wouthuysen-Field coupling became efficient.

However, the details of radiative transfer radically change these expectations. In a static medium, the energy exchange *must* vanish in equilibrium even though scattering continues at nearly the same rate. Scattering induces an asymmetric absorption feature near ν_α (Figure 12.6) whose shape depends on the combined effects of atomic recoils and the scattering diffusivity. In equilibrium, the latter exactly counterbalances the former.

In an expanding universe without scattering, the absorption feature would redshift away; thus, the equilibrium energy exchange rate is simply that required to maintain the feature in place. For photons redshifting into resonance, the absorption trough has total energy

$$\Delta u_\alpha = (4\pi/c)\int (J_\infty - J_\nu) h\nu d\nu, \tag{12.34}$$

insert small space / PE

where J_∞ is the input spectrum (thus, the integration extends over the dip in Figure 12.6). The radiation background loses $\epsilon_\alpha = H\Delta u_\alpha$ per unit time through redshifting; this energy goes into heating the gas. Relative to adiabatic cooling by the Hubble expansion, the fractional heating amplitude is[14]

$$\frac{2}{3}\frac{\epsilon_\alpha}{k_B T_K n_{\rm H} H(z)} = \frac{8\pi}{3}\frac{h\nu_\alpha}{k_B T_K}\frac{J_\infty \Delta\nu_D}{c n_{\rm H}}\int_{-\infty}^{\infty} dx\, \delta_J(x) \tag{12.35}$$

$$\approx \frac{0.80}{T_K^{4/3}}\frac{x_\alpha}{S_\alpha}\left(\frac{10}{1+z}\right), \tag{12.36}$$

. (period) / AA

Here we have evaluated the integral for the continuum photons that redshift into the Lyman-α resonance; the "injected" photons actually cool the gas slightly. The net energy exchange when Wouthuysen-Field coupling becomes important (at $x_\alpha \sim S_\alpha$) is therefore just a fraction of a degree, and in practice, gas heating through Lyman-α scattering is usually unimportant.

The reason for the inefficiency of heating is that the scattering diffusivity acts to cancel the effects of recoil. From Figure 12.6, it is obvious that the background spectrum is weaker on the blue side of the line than on the red. Scattering tends to return the photon toward line center, with the extra energy

[viii] Moreover, the ionized regions do not significantly recombine unless somehow the source emissivity declines dramatically. The best example is a highly luminous quasar that ionizes a large region around itself and then shuts off shortly thereafter.

deposited in or extracted from the gas. Because more scattering occurs on the red side, this tends to transfer energy from the gas back to the photons, canceling the recoil exchange.

Thus, IGM heating is likely dominated by X-rays—whether from Population III stars, supernova remnants, stellar-mass black holes, or quasars. We have already seen (equation 9.65) that X-rays from a "reasonable" quasar population can have a dramatic effect on the IGM, but even the weaker X-ray emissivity of stellar-mass black holes can also be significant.

A simple, but plausible, way to parameterize this emissivity is with the local correlation between the star formation rate (SFR) and the X-ray luminosity in the photon energy band of 0.5–8 keV,[15]

$$L_X = 3 \times 10^{39} f_X \left(\frac{\mathrm{SFR}}{M_\odot\,\mathrm{yr}^{-1}} \right) \mathrm{erg\,s}^{-1}, \tag{12.37}$$

where f_X is an unknown renormalization factor appropriate for high redshifts. We can only speculate as to the accuracy of this correlation at higher redshifts. Certainly, the scaling is appropriate so long as recently formed remnants dominate, but f_X likely evolves with redshift.

The X-ray emission has two major sources. The first is inverse-Compton scattering off relativistic electrons accelerated in supernovae. In the nearby Universe, only powerful starbursts have strong enough radiation fields for this mechanism to be significant; however, at high redshifts it probably plays an increasingly important role, because the CMB energy density $u_\gamma \propto (1+z)^4$. Assuming that ~5% of the supernova energy is released in this form yields $f_X \sim 5$ if $\sim 10^{51}$ erg is released in supernovae per $100\,M_\odot$ in star formation. The second class of sources, which dominate in locally observed galaxies, are high-mass X-ray binaries, in which material from a massive main-sequence star accretes onto a compact neighbor. Such systems were born as soon as the first massive stars died, only a few million years after star formation commenced. So they certainly ought to exist in high-redshift galaxies, although their abundance depends on the metallicity and stellar initial mass function. To the extent that massive stars are more abundant at high redshifts (see the discussions of the IMF in chapter 5), we would expect such binaries also to be more abundant, which is consistent with some observational hints of evolution in this relation toward higher redshifts.

Regardless of the details of the sources, the heating rate and temperature profile around each source can be computed following the methods in §9.8.2. Note that, unlike the Wouthuysen-Field background, the IGM temperature depends not on the *instantaneous* emissivity of sources but on the accumulated emissivity over the entire history of structure formation. Thus, the IGM temperature structure is more complicated to compute, although the same basic picture—built from the effects of each source halo—applies.

Figure 12.8 shows the temperature histories and power spectra of T_K for two models in which the heating is due to star-forming galaxies. In the left panels, the thick lines take $f_X = 10$ and standard Population II stars, forming with an

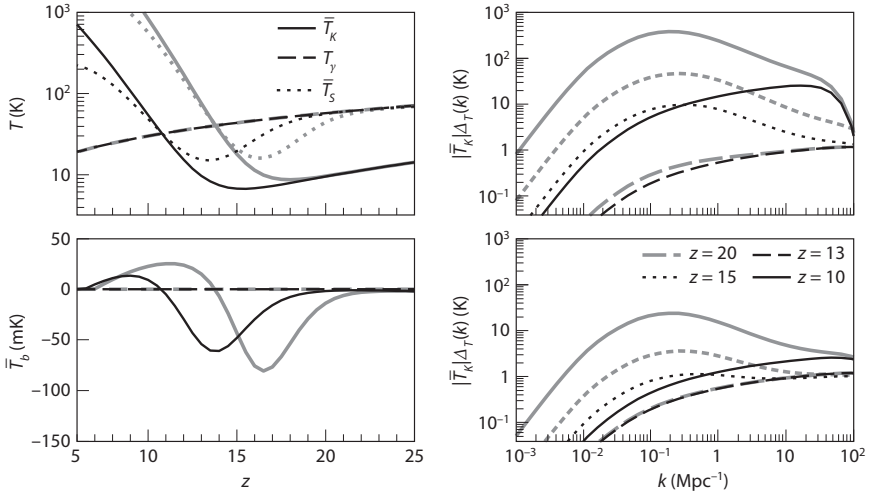


Figure 12.8 *Left:* Thermal history of two models of X-ray heating, Wouthuysen-Field coupling, and reionization. In the top panel, the solid, dashed, and dotted lines show T_K, T_γ, and T_S, respectively. The ~~thick~~ and ~~thin~~ lines take Population II and III star formation properties, respectively, (see text for details). The bottom panel shows the mean spin-flip brightness temperature T_b in the same two models. *Right:* Power spectrum of temperature fluctuations in the same models, from $z = 20$ (where the mean temperature is nearly that of an adiabatically cooling IGM) to $z = 10$ (where $T_K \gg T_\gamma$). The upper and lower panels are for the Population II and Population III models shown at left, respectively. The peak at late times corresponds to the typical mean free path of X-ray photons. The thin curves show the fluctuations for uniform heating for comparison (in this case the fluctuations arise from variations in the expansion cooling rate). Pritchard, J. R., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **376**, 1680 (2007). Copyright 2007 by the Royal Astronomical Society.

light / dark / AA

dark / AA

efficiency $f_\star = 0.1$ in halos with $T_{\rm vir} > 10^4$ K; the thin lines are identical but take $f_\star = 0.01$ and use very massive Population III stars to determine the UV properties. Note that even with this relatively modest heating rate, heating begins at $z \sim 15$, and the IGM temperature surpasses T_γ shortly thereafter. The right panels show the corresponding temperature power spectra; the top and bottom panels are for the Population III and Population II models, respectively. In absolute terms, the temperature fluctuations begin quite modestly; at $z = 20$ they are driven primarily by variations in the adiabatic cooling rate with IGM density. By $z = 15$, the fractional fluctuations are ∼20%—which will translate into large 21-cm fluctuations. The absolute amplitude of the fluctuations continues to increase at lower redshifts, but the fractional fluctuations decrease as more sources appear. Moreover, in the limit $T_K \gg T_\gamma$, the 21-cm brightness temperature is independent of T_K, so these fluctuations are unimportant.

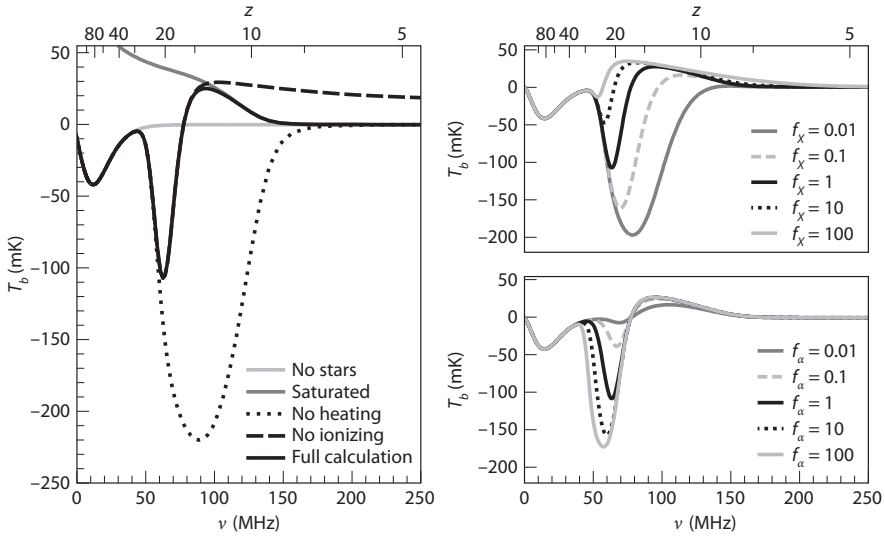


Figure 12.9 Monopole of the spin-flip brightness temperature (the so-called global 21-cm signal) in several models of early structure formation (see *Color Plate 27* for a color version of this figure). *Left:* Major variations around our fiducial model (solid curve with several turning points), as indicated. Each curve either eliminates a physical process (like heating or ionization) or maximizes it. *Right:* Suites of models in which the Lyman-α (lower panel) and X-ray heating (upper panel) efficiencies are varied by a factor of 10^4. Pritchard, J. R., & Loeb, A., *Phys. Rev.* **D82**, 023006 (2010). Copyright 2010 by the American Physical Society.

12.4 The Monopole of the Brightness Temperature

We are now in a position to compute the time evolution of the brightness temperature T_b in some simple models. We begin in this section with the monopole, or sky-averaged brightness, as a function of frequency. Figure 12.9 shows the results (as functions of both redshift and observed frequency) for a set of models, illustrating the wide range of possible histories. At left we show some highly simplified models. The solid curve with several turning points is our fiducial model, in which we take $f_X = 1$ and standard Population II stars, forming with an efficiency $f_\star = 0.1$ in halos with $T_{\rm vir} > 10^4$ K. The other two solid lines show histories with no star formation (flat below $z \sim 30$) and with a hot, fully coupled IGM (descending from large T_b). The dashed curve shows a history in which reionization does not occur, and the dotted curve shows a history in which heating is turned off.

The right panels take somewhat more sophisticated models, in which the X-ray heating efficiency (via f_X, see equation 12.37) and Lyman-α intensity (via a parameter f_α, defined so that the intensity from each galaxy is f_α times that in the fiducial model) are varied by factors of 10^4.

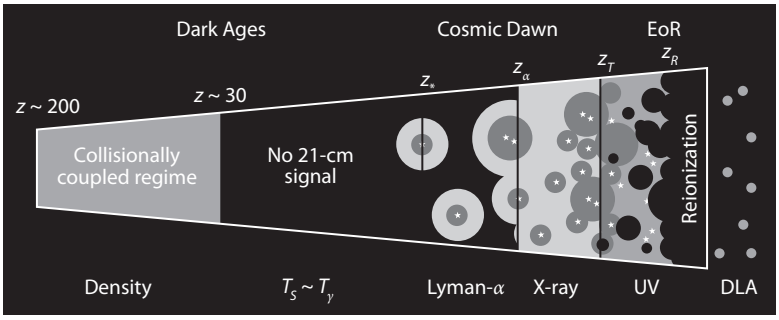


Figure 12.10 Cartoon of the different phases of the 21-cm signal (see *Color Plate 28* for a color version of this figure). The signal transitions from an early phase of collisional coupling to a later phase of Lyman-α coupling through a short period where there is little signal. Fluctuations after this phase are dominated successively by spatial variation in the Lyman-α, X-ray, and ionizing UV radiation backgrounds. After reionization is complete, there is a residual signal from neutral hydrogen in galaxies. Pritchard, J. R., & Loeb A., *Reports on Progress in Physics*, in press (2012), http://arxiv.org/pdf/1109.6012.

These different models are essentially cartoons, but they illustrate several important points about the 21-cm background. The most important is the presence of five critical points in the spin-flip background, separating the eras shown in Figure 12.10.

1. The first, at $z \sim 80$, occurs long before star formation becomes significant during the collisionally coupled regime. This point reflects the decreasing effectiveness of collisional coupling and occurs roughly when the density $\delta_{\rm coll}$ falls below unity (see equation 12.15), at which point $T_S \rightarrow T_\gamma$, and the IGM signal fades. This transition is well specified by atomic physics and the standard cosmology, at least in the absence of any exotic dark sector processes that may input energy into the IGM at $z > 50$. This signal therefore provides a clear probe of cosmology, at least in principle.
2. The remaining transition points are determined by the properties of luminous sources, so their timing is much more uncertain. In our fiducial model, the next crucial event is the formation of the first stars (at $z \sim 30$), which flood the Universe with Lyman-α photons and so reignite the 21-cm background. Interestingly, the timing of this transition is relatively independent of the luminosity of these sources, because (at least in this model) the massive halos that host these sources are so far out on the exponential tail of the mass function that their luminosity is primarily determined by the rate of halo collapse. Thus, this turning point primarily constrains the characteristic mass of the first galaxies.
3. Next (usually), is the minimum in T_b, which occurs just before IGM heating begins to become significant and is determined by f_X. However, if f_X is very large compared with f_α, this heating transition can precede strong

coupling. In simple models like we use here, in which both the X-ray and UV luminosities trace $f_{\rm coll}$, the net X-ray heat input ΔT_c when $x_\alpha = 1$ is

$$\frac{\Delta T_c}{T_\gamma} \sim 0.08 f_X \left(\frac{f_{X,h}}{0.2}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{9,690}{N_\alpha}\frac{1}{S_\alpha}\right)\left(\frac{20}{1+z}\right)^3, \qquad (12.38)$$

close up / PE

where $\Delta f_{\rm coll} \sim f_{\rm coll}$ is the effective collapse fraction appearing in the integrals of equation (12.30), and $f_{X,h}$ is the fraction of the X-ray energy that goes into heating (see §9.8.2). Note that ΔT_c is independent of $f_\star$, because we have assumed that both the coupling and heating rates are proportional to the star formation rate. Clearly, for our fiducial (Population II) parameters the onset of Wouthuysen-Field coupling precedes the point at which heating begins, which is ultimately the reason for the strong absorption in our fiducial model.

4. The fourth turning point is at the maximum of T_b. In the fiducial model, this marks the point at which $T_K \gg T_\gamma$, so that the temperature part of equation (12.8) saturates. From that time forward, the only factors affecting the monopole are the redshift and the ionized fraction, so the signal starts to decrease rapidly once reionization begins in earnest. Most likely, this happens *after* coupling is already strong and heating is significant. Again, in the simple models used here the ionized fraction when $x_\alpha = 1$ is given by

$$\bar{x}_{i,c} \sim 0.05 \left(\frac{f_{\rm esc}}{1+\bar{n}_{\rm rec}}\frac{N_{\rm ion}}{N_\alpha}\frac{f_{\rm coll}}{\Delta f_{\rm coll}}\frac{1}{S_\alpha}\right)\left(\frac{20}{1+z}\right)^2, \qquad (12.39)$$

where $\bar{n}_{\rm rec}$ is the mean number of recombinations per baryon. For Population II stars with a normal IMF, $N_{\rm ion}/N_\alpha \approx 0.4$; thus, even in the worst case, where $f_{\rm esc} = 1$ and $\bar{n}_{\rm rec} = 0$, coupling will become efficient during the initial stages of reionization. However, very massive Population III stars have much harder spectra, with $N_{\rm ion}/N_\alpha \approx 7$. In principle, it is therefore possible for Population III stars to reionize the universe *before* $x_\alpha = 1$, although this is rather unlikely given their fragility (see chapter 6). It is less clear whether the IGM will appear in absorption or emission during reionization. We find

$$\frac{\Delta T}{T_\gamma} \sim \left(\frac{\bar{x}_i}{0.025}\right)\left(f_X \frac{f_{X,h}}{f_{\rm esc}}\frac{4800}{N_{\rm ion}}\frac{10}{1+z}\right)(1+\bar{n}_{\rm rec}) \qquad (12.40)$$

for the heat input ΔT as a function of $\bar{x}_i$. Thus, provided $f_X > 1$, the IGM will be much warmer than the CMB during the bulk of reionization. But the right panel of Figure 12.9 shows that this is by no means assured.

5. The monopole signal (nearly) vanishes when reionization is complete; the residual brightness is due to gas that is self-shielded from the metagalactic ionizing background (and hence primarily lies inside galaxies, since the LLSs still have small ionized fractions).

Several efforts to observe this monopole signal are underway, including the Cosmological Reionization Experiment (CoRE) and the Experiment to Detect

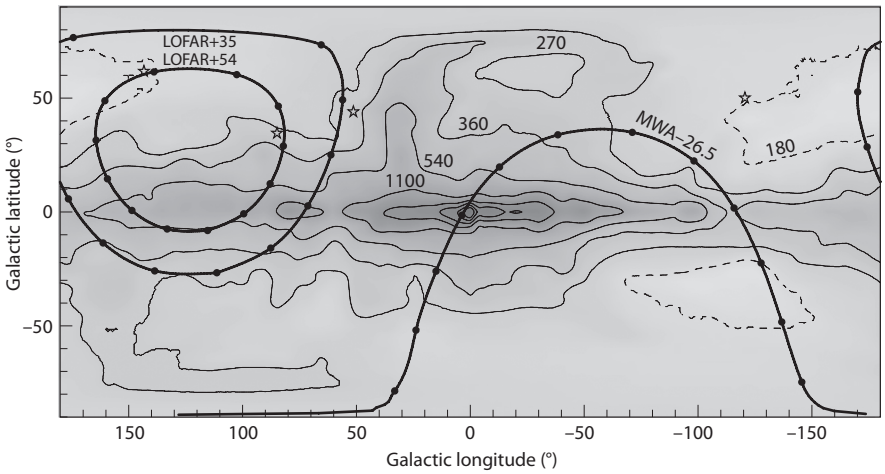


Figure 12.11 Brightness temperature of the radio sky at 150 MHz in Galactic coordinates. Contours are drawn at 180 (dashed), 270, 360, 540, 1,100, 2,200, 3,300, 4,400, and 5,500 K. Heavy lines indicate constant declinations: −26.5°, +35°, and +54° with dots to mark 2-hour intervals of time (these are ideal for two other existing experiments, the Murchison Widefield Array, or MWA; and LOFAR, or Low Frequency Array). Star symbols indicate the coordinates of four bright $z>6.2$ quasars. Landecker, T. L., & Weilebinski, R., *Proc. Astron. Soc. Aust.* **1**, 210 (1969); Furlanetto, S. R., Oh, S. P., & Briggs, F. H., *Physics Reports* **433**, 181 (2006). Copyright 2006 by Elsevier.

the Global Epoch of Reionization Signal (EDGES)[16]. The wide range of histories shown in Figure 12.9 illustrates how powerful such observations would be.

Because global experiments aim to detect an all-sky signal, single-dish measurements (even with a modest-sized telescope) can easily reach the required millikelvin sensitivity. However, the much stronger synchrotron foregrounds from our Galaxy nevertheless make such observations extremely difficult: they have $T_{\rm sky} > 200$–10^4 K over the relevant frequencies (see the map in Figure 12.11). The fundamental strategy for extracting the cosmological signal relies on the expected spectral smoothness of the foregrounds (which primarily have power-law synchrotron spectra), in contrast with the nontrivial structure of the 21-cm background. Nevertheless, isolating the high-redshift component will be a challenge requiring extremely accurate calibration over a wide frequency range and, most likely, sharp localized features in $T_b(z)$ that can be distinguished from smoother foreground features.

Current estimates show that rapid reionization histories that span a redshift range $\Delta z < 2$ can be constrained, provided that local foregrounds can be well modeled.[17] Observations in the frequency range 50–100 MHz can potentially constrain the Lyman-α and X-ray emissivity of the first stars and black holes: even though the foregrounds are significantly worse at these lower frequencies,

Reprinted from *Phys. Rep.*, 433, 4-6, Furlanetto, S.R., S.P. Oh, & F.H. Briggs, "Cosmology at low frequencies: The 21cm transition and the high-redshift Universe," 181-301, Copyright 2006, with permission from Elsevier. Based on data from Landecker, T.L. & Wielebinski, R., Aust. J. Phys. Astrophys. Supp. 16 (1970).

the strong absorption signal present in many models may be easier to observe than the gently varying reionization signal. However, it may be necessary to perform such observations from space, to avoid systematics from terrestrial interference and the ionosphere (in fact, the best observing environment is the far side of the moon, where the moon itself blocks any radio signals from Earth).

12.5 Statistical Fluctuations in the Spin-Flip Background

While the 21-cm monopole contains a great deal of information about the mean evolution of the sources, every component in equation (12.8) can also fluctuate significantly. For the density field this is obvious: the evolving cosmic web imprints growing density fluctuations on the matter distribution. For some of the other aspects, the luminous sources cause 21-cm fluctuations. Ionized gas is organized into discrete H II regions (at least in the most plausible models), and the Lyman-α background and X-ray heating are also concentrated around galaxies. The single greatest advantage of the 21-cm line is that it allows us to separate this fluctuating component both on the sky and in frequency (and hence cosmic time). Thus, we can study the sources and their effects on the IGM in detail. It is the promise of these "tomographic" observations that makes the 21-cm line such a singularly attractive probe.

Observing the 21-cm fluctuations has one practical advantage as well. The difficulty of extracting the global evolution from the enormously bright foregrounds shown in Figure 12.11 lies in its relatively slow variation with frequency. On the small scales relevant to fluctuations in the signal, the gradients increase dramatically: for example, at the edge of an H II region T_b drops by $\sim$20 mK essentially instantaneously. As a result, separating fluctuations from the smoothly varying astronomical foregrounds may be much easier. Unfortunately, constructing detailed images will remain extremely difficult because of their extraordinary faintness; telescope noise is comparable to or exceeds the signal except on rather large scales. Thus, a great deal of attention has recently focused on using statistical quantities readily extractable from low signal-to-noise maps to constrain the IGM properties. This technique is motivated in part by the success of CMB measurements and galaxy surveys at constraining cosmological parameters through the power spectrum. In our case, although any number of statistical quantities may be useful (especially during reionization, when the fluctuations are highly non-Gaussian), we take the power spectrum as our primary analysis tool.

We first define the fractional perturbation to the brightness temperature, $\delta_{21}(\mathbf{x}) \equiv [T_b(\mathbf{x}) - \bar{T}_b]/\bar{T}_b$, a zero-mean random field. We are interested in its Fourier transform $\tilde{\delta}_{21}(\mathbf{k})$. Its power spectrum is defined to be

$$\left\langle \tilde{\delta}_{21}(\mathbf{k}_1)\, \tilde{\delta}_{21}(\mathbf{k}_2) \right\rangle \equiv (2\pi)^3 \delta_D(\mathbf{k}_1 - \mathbf{k}_2)\, P_{21}(\mathbf{k}_1), \tag{12.41}$$

where $\delta_D(x)$ is the Dirac delta function, and the angular brackets denote an ensemble average. Power spectra for other random fields (such as the fractional

overdensity δ, the ionized fraction, etc.), or cross–power spectra between two different fields, can be defined in an analogous fashion.

As is obvious from equations (12.8) and (12.10), the brightness temperature depends on a number of input parameters. Expanding those equations to linear order in each of the perturbations, we can write

$$\delta_{21} = \beta\delta_b + \beta_x\delta_x + \beta_\alpha\delta_\alpha + \beta_T\delta_T - \delta_{\partial v}, \tag{12.42}$$

where each δ_i describes the fractional variation in a particular quantity: δ_b for the baryonic density (for which we use the matter density, though the baryonic density is smoother on very small scales owing to pressure smoothing), δ_α for the Lyman-α coupling coefficient x_α, δ_x for the neutral fraction (note that using the ionized fraction would cause a sign change), δ_T for T_K, and $\delta_{\partial v}$ for the line-of-sight peculiar velocity gradient. The expansion coefficients β_i are

$$\beta = 1 + \frac{x_c}{x_{\rm tot}(1+x_{\rm tot})}, \tag{12.43}$$

$$\beta_x = 1 + \frac{x_c^{\rm HH} - x_c^{\rm eH}}{x_{\rm tot}(1+x_{\rm tot})} \tag{12.44}$$

$$\beta_\alpha = \frac{x_\alpha}{x_{\rm tot}(1+x_{\rm tot})}, \tag{12.45}$$

$$\beta_T = \frac{T_\gamma}{T_K - T_\gamma} + \frac{1}{x_{\rm tot}(1+x_{\rm tot})}\left(x_c^{\rm eH}\frac{d\ln\kappa_{10}^{\rm eH}}{d\ln T_K} + x_c^{\rm HH}\frac{d\ln\kappa_{10}^{\rm HH}}{d\ln T_K}\right), \tag{12.46}$$

where $x_{\rm tot} \equiv x_c + x_\alpha$, and we have split the collisional term into the dominant H–e$^-$ and H–H components ($x_c^{\rm eH}$ and $x_c^{\rm HH}$, respectively) where necessary. Here we have assumed $T_c = T_K$ throughout; this is reasonable in most cases but, if not, the expressions become much more complicated. By linearity, the Fourier transform $\tilde{\delta}_{21}$ can be written in a similar fashion. (For now we ignore the velocity term; see §12.5.1)

Each of these expressions has a simple physical interpretation. For β, the first term describes the increased matter content, and the second describes the increased collisional coupling efficiency in dense gas. For β_x, the two terms describe direct fluctuations in the ionized fraction and the effects of the increased electron density on x_c. (The latter is important only in partially ionized regions; 21-cm emission is negligible in H II regions, of course.) β_α simply measures the fractional contribution of the Wouthuysen-Field effect to the coupling. The first term in β_T parameterizes the speed at which the spin temperature responds to fluctuations in T_K, while the others include the explicit temperature dependence of the collision rates. Note that all these terms, with the crucial exception of $\delta_{\partial v}$, are isotropic; we discuss this latter effect in the next section.

For context, Figure 12.12 shows how these expansion coefficients evolve in a typical structure formation model (similar to those described in the previous section). The density coefficient β increases with time until $z \sim 20$ before abruptly falling to unity. At $z > 20$, collisions are only marginally important, so the extra collisional coupling imparted by an increased density has a relatively large effect; at lower redshifts, collisional coupling is negligible compared

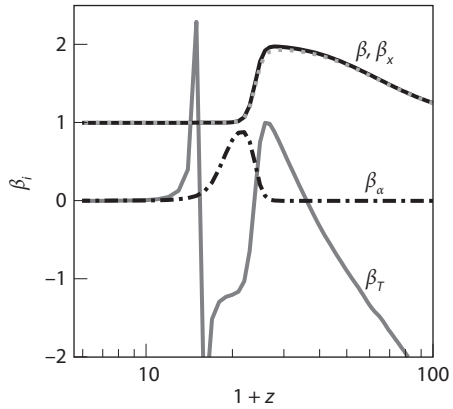


Figure 12.12 Redshift dependence of perturbative expansion coefficients in a fiducial model similar to that of Figure 12.9. We show β (solid curve), β_x (dotted curve), β_α (dot-dashed curve), and β_T (~~dashed~~ curve). Note that the singularity in β_T at $z = 17$ is artificial in that it does not actually appear in the fluctuation amplitude. Furlanetto, S. R., Oh, S. P., & Briggs, F. H., *Physics Reports* **433**, 181 (2006). Copyright 2006 by Elsevier.

light / AA

with the Wouthuysen-Field effect, so the second term in equation (12.43) vanishes. β_x behaves nearly identically, because (outside H II regions) the ionized fraction remains small. Fluctuations in the Lyman-α background are important only over a limited redshift range (where $x_\alpha \sim 1$, or the coupling is marginal); at lower redshifts, all the gas is strongly coupled, so fluctuations in the background are unimportant. The temperature coefficient has the most complicated dependence, because it depends on the mix of Compton heating and collisional coupling. Note that the apparent singularity occurs where $T_K = T_\gamma$; it is not physical, because $\bar{T}_b$ also vanishes at the same point. At lower redshifts, $T_K \gg T_\gamma$, and the emission saturates: $\beta_T \to 0$.

Based on equation (12.41), the power spectrum contains all possible terms of the form $P_{\delta_i \delta_j}$; some or all could be relevant in any given situation. Of course, in most instances the various δ_i will be correlated in some way; statistical 21-cm observations ideally hope to measure these separate quantities. We have already included some of the obvious correlations in equations (12.43)–(12.46), such as the variation of the collision rate with the ionized fraction. But we have left others implicit: for example, overdense regions are ionized first in most reionization models. A more subtle example is the relation of δ_α to the other quantities: as we saw in §12.2.2, it depends on the radiation spectrum and hence on density, neutral fraction, and temperature in addition to the background flux.

In all these expansions, one must bear in mind that δ_x is *always* of the order of unity if the ionization field is built from H II regions, because in that case $x_i = 0$ or 1, and terms such as $\delta\delta_x$ are, in fact, *first* order and must be retained in detailed calculations.

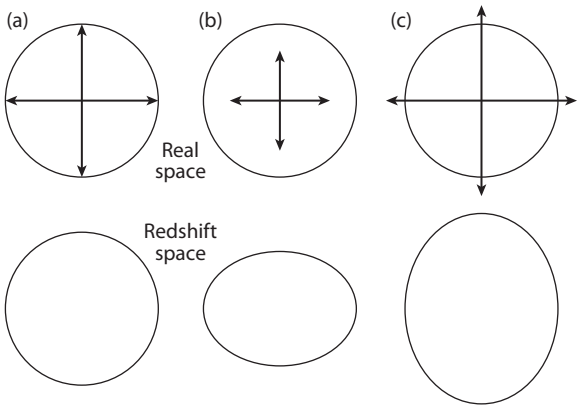


Figure 12.13 Cartoon of redshift space distortions in the linear regime. If an observation uses redshift as a proxy for distance, it is subject to confusion if velocities do not precisely follow the mean Hubble expansion. In *(a)*, we imagine observing a region at the mean density that follows the Hubble flow (top row). In that case, the inferred radial width of the region will be precisely correct (bottom row). However, if the region is overdense, so that the local expansion rate is smaller than the Hubble flow, the inferred radial width will be *smaller* than its true width, so the observed density will be even larger (column *b*). Finally, if the region is underdense (and so expanding faster than average, it will appear even more underdense in redshift space (column *c*).

12.5.1 Redshift-Space Distortions

In general, we expect the fluctuations in density, ionization fraction, Lyman-α flux, and temperature to be statistically isotropic, because the physical processes responsible for them have no preferred direction [e.g., $\delta(\mathbf{k}) = \delta(k)$].[ix] However, peculiar velocity gradients introduce anisotropic distortions. Bulk flows on large scales, and, in particular, infall onto massive structures, compress the signal in redshift space (the so-called Kaiser effect),[18] enhancing the apparent clustering amplitude, as illustrated in Figure 12.13. On small scales, random motions in virialized regions create elongation in redshift space (the "finger of God" effect), reducing the apparent clustering amplitude.[x]

We start by labeling the coordinates in redshift space by $\mathbf{s}$. Working for simplicity under the assumption that the survey volume has a small radial depth (so that the Hubble parameter H can be considered constant throughout the

[ix] Actually, this assumption can break down on extremely large scales, because then the growth of structure with redshift becomes important. Fortunately, the 21-cm brightness field contains only rapidly evolving features on such large scales near the tail end of reionization. The evolution is generally not important on the scales accessible to observations.

[x] In most applications, these tend to wash out fluctuations in redshift space. Fortunately, this effect is negligible for the spin-flip background because the vast majority of the gas lies outside massive virialized structures (and gas inside such halos is almost always inside ionized regions anyway).

volume), we can write for the relation of these coordinates to real space $\mathbf{r}$

$$\mathbf{s}(\mathbf{r}) = \mathbf{r} + \frac{U(\mathbf{r})}{H}, \tag{12.47}$$

where $U(\mathbf{r}) = \mathbf{v} \cdot \hat{\mathbf{x}}$ is the radial component of the peculiar velocity.

Next, we consider a set of particles with number density $n(\mathbf{r})$ that are biased with respect to the dark matter by a factor b. Number conservation demands that the fractional overdensity in redshift space be related to that in real space via $[1 + \delta_s(\mathbf{s})]d^3\mathbf{s} = [1 + \delta(\mathbf{r})]d^3\mathbf{r}$. The Jacobian of the transformation is

$$d^3\mathbf{s} = d^3\mathbf{r}\left[1 + \frac{U(\mathbf{r})}{r}\right]^2\left[1 + \frac{dU(\mathbf{r})}{dr}\right], \tag{12.48}$$

because only the radial component of the volume element, $r^2 dr$, changes from real to redshift space. The density observed in redshift space increases if the peculiar velocity gradient is smaller than the Hubble flow and decreases otherwise. Thus, assuming $|U(r)| \ll Hr$,

$$\delta_s(\mathbf{r}) = \delta(\mathbf{r}) - \left(\frac{d}{dr} + \frac{2}{r}\right)\frac{U(r)}{H}. \tag{12.49}$$

Conveniently, the peculiar velocity field itself is a function of the dark matter density field, as described by equation (~~2.14~~).

2.13 / renumbering

To see which of these corrections is more important, consider a plane-wave perturbation, $U \propto e^{i\mathbf{k}\cdot\mathbf{r}}$. Then, the derivative term is $\sim kU/H_0$, while the last term is $\sim U/H_0 r$. But r is the median distance to the survey volume, and k corresponds to a mode entirely contained inside it. For all but the largest surveys, we must therefore have $kr \gg 1$, and we may neglect the last term. If we further make the small-angle approximation, so that $\hat{\mathbf{x}}$ is also approximately a constant over the survey volume, we can take the Fourier transform of equation (12.49) and find

$$\delta_s(\mathbf{k}) = \delta(\mathbf{k})[1 + \beta\mu_{\mathbf{k}}^2] \tag{12.50}$$

where $\mu_{\mathbf{k}} = \hat{\mathbf{k}} \cdot \hat{\mathbf{x}}$ is the cosine of the angle between the wave vector and the line of sight, and we have used (see equation ~~2.14~~)

$$U(r) = \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}}[-i\beta\delta(\mathbf{k})]\frac{\hat{\mathbf{k}} \cdot \hat{\mathbf{x}}}{k}. \tag{12.51}$$

Here $\beta = f(\Omega_m)/b$ corrects for a possible bias between the tracers we are studying and the growth rate of dark matter perturbations, and $f(\Omega_m) \approx \Omega_m^{0.6}(z)$. For the case of 21-cm fluctuations in the IGM gas, the bias factor is very close to unity except below the Jeans filtering scale. Moreover, at high redshifts, when the universe is matter dominated, $f \approx 1$.

The redshift-space distortions therefore provide an anisotropic *amplification* to the background signal. The anisotropy occurs because only modes along the line of sight are affected, as illustrated in Figure 12.13. To understand the amplification, consider a spherical overdense region. Its excess gravitational force

causes it to recollapse. Along the radial direction, the collapse *decreases* the velocity width of the object relative to the Hubble flow (at least in linear theory), compressing the overdensity in redshift space. Similarly, a spherical underdensity expands faster than average, causing it to appear elongated in the radial direction. Averaged over all modes, these distortions amplify the signal by a factor $\approx \left\langle (1+\mu^2)^2 \right\rangle \approx 1.87$.

However, the anisotropies are actually even more helpful in that they provide angular structure to the signal, which may allow us to separate the many contributions to the total power spectrum. Schematically, brightness temperature fluctuations in Fourier space have the form

$$\delta_{21} = \mu^2 \beta \delta + \delta_{\rm iso} \tag{12.52}$$

where we have collected all the statistically isotropic terms in equation (12.42) into $\delta_{\rm iso}$. Neglecting "second-order" terms (see later discussion) and setting $\beta = 1$, we can therefore write the total power spectrum as [19]

$$P_{21}(\mathbf{k}) = \mu^4 P_{\delta\delta} + 2\mu^2 P_{\delta_{\rm iso}\delta} + P_{\delta_{\rm iso}\delta_{\rm iso}}. \tag{12.53}$$

By separately measuring these three angular components (which requires, in principle, estimates at just a few values of μ), we can isolate the contribution from density fluctuations $P_{\delta\delta}$. This would not have been possible without peculiar velocity flows: comparison with equation (12.42) shows that, in the most general case, $P_{\delta_{\rm iso}\delta}$ and $P_{\delta_{\rm iso}\delta_{\rm iso}}$ contain several different power spectra, including those of the density, neutral fraction, and spin temperature as well as their cross–power spectra.

Disentangling these other components is more difficult, since there are several remaining power spectra to be determined from the two measured quantities $P_{\delta_{\rm iso}\delta}(k)$ and $P_{\delta_{\rm iso}\delta_{\rm iso}}(k)$. Fortunately, in many regimes one or more of the terms can be neglected. For example, during the earliest stages of reionization (when δ_x is negligible), one might be able to measure the power spectrum of spin temperature fluctuations as well as its correlations with density. At late times (when $T_S \gg T_\gamma$, and T_b becomes independent of T_S), one might likewise ignore spin temperature fluctuations and measure the ionization fraction fluctuations $P_{\delta x}$ and P_{xx}.

An additional difficulty originates from the correlations of "second-order" terms in the perturbation expansion, such as $\delta\delta_x$, that produce four-point terms in the power spectrum. As mentioned previously, δ_x is not necessarily a small parameter, so these terms can be substantial, and in practice they can produce terms with nontrivial μ dependence, especially during reionization. The presence of these terms make attempts to separate the μ^n powers during reionization more difficult; the prospects are much better before δ_x becomes important.

Another important caveat to recovering redshift-space distortions is that they require a high signal-to-noise measurement of the angular structure of the signal. Unfortunately, the noise is anisotropic: radio foregrounds have much more power across the sky than in the line-of-sight direction. (Indeed, this very feature is crucial to foreground-removal algorithms.) Moreover, it is much easier to probe small physical scales in the frequency direction than across the

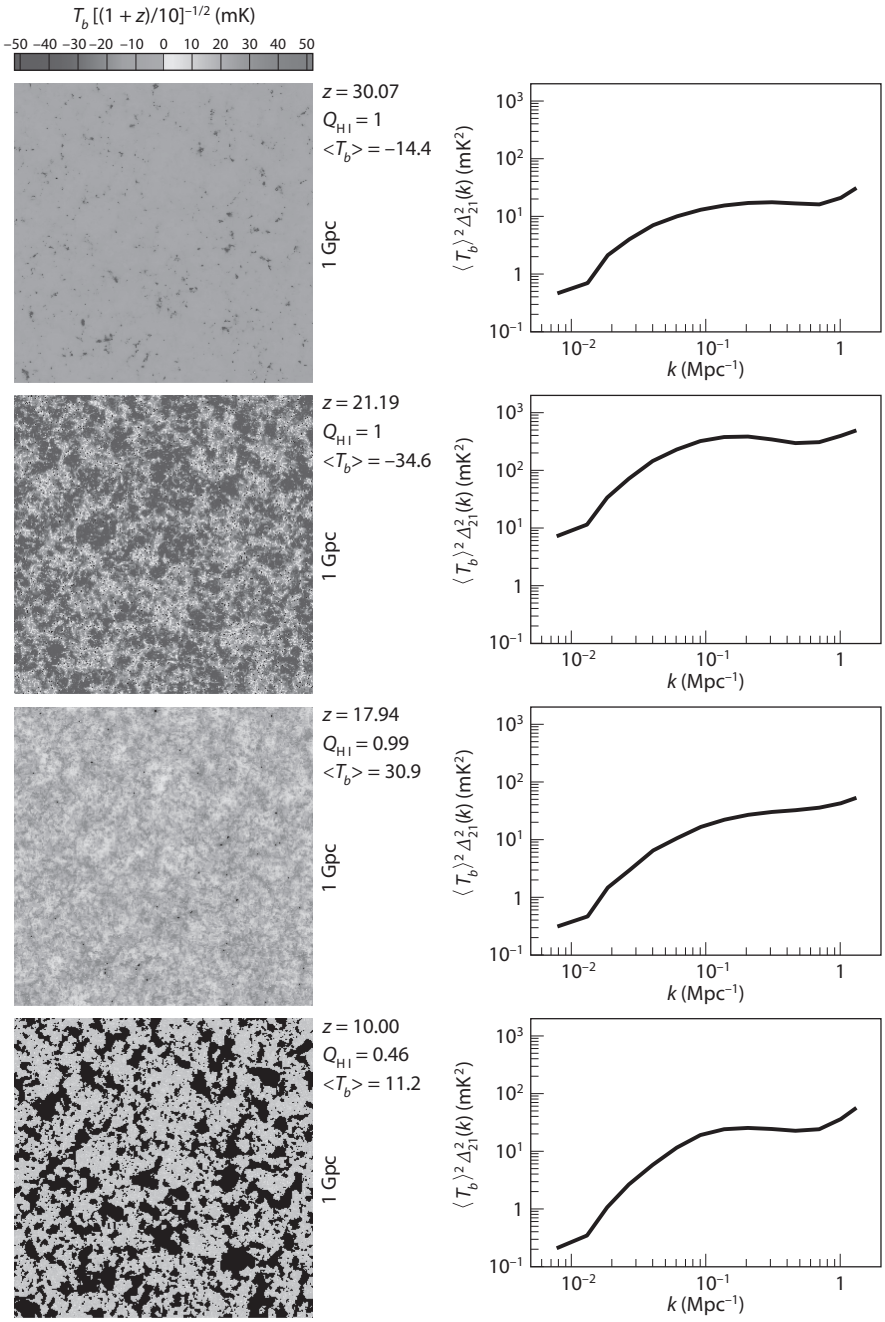


Figure 12.14 Slices through a "seminumerical" simulation (*left*), and the corresponding spherically averaged power spectra (*right*), for a model of the spin-flip background at $z = 30.1, 21.2, 17.9, 10.0$ (*top to bottom*); see *Color Plate 29* for a color version of this figure. The slices were chosen to highlight various

angular dimensions. As a result, taking advantage of this "separation of angular powers" will be difficult.

12.5.2 Other Statistical Measures

So far we have focused our discussion on the three-dimensional power spectrum, which is familiar to most cosmologists and provides a reasonable description of the spin-flip background during most of its evolution. In fact the power spectrum is a complete statistical description of any purely Gaussian random field (whose only parameters are, by definition, the mean and variance as a function of spatial scale). Inflation predicts that the initial matter density field is nearly Gaussian, making the power spectrum a powerful tool in cosmology.

However, nonlinear evolution—and the radiation fields from luminous sources—spoil this simple statistical description for the 21-cm fluctuations, especially when ionized bubbles become prevalent late in reionization. It is easy to see that a Gaussian probability distribution can no longer adequately describe the 21-cm field during these periods: at infinite spatial resolution, the signal is either nearly zero (in an ionized bubble) or ∼20 mK (in the neutral gas, where there is still some variation owing to the density field and possibly T_S). This bivariate distribution is a strong signature of ionized bubbles and would provide a powerful test of the morphology of reionization; unfortunately, in experiments where the Gaussian noise per pixel is larger than ∼20 mK, this kind of distribution may be difficult to detect, especially in the presence of complex astrophysical foregrounds.

Other statistical measures, such as higher-order correlations, may also offer additional information and are the subject of ongoing research in the community.

12.6 Spin-Flip Fluctuations During the Cosmic Dawn

Figure 12.14 shows several snapshots of a "seminumerical" computer simulation (see §9.6.2) of the spin-flip background during the important stages outlined in our discussion of the monopole signal of §12.4, including both

Figure 12.14 (Continued) epochs in the cosmic 21-cm signal (from top to bottom): the onset of Lyman-α pumping (here the dark regions show the cold gas around the first galaxies); the onset of X-ray heating (here the dark regions are cold gas, while the compact spots represent hot gas around the first black holes); the completion of X-ray heating (where all the gas is hot); and the midpoint of reionization (where black regions are ionized bubbles). All slices are 1 Gpc on a side and 3.3 Mpc deep. Mesinger, A., Furlanetto, S. R., & Cen, R., *Mon. Not. R. Astron. Soc.* **411**, 955 (2011). Copyright 2011 by the Royal Astronomical Society.

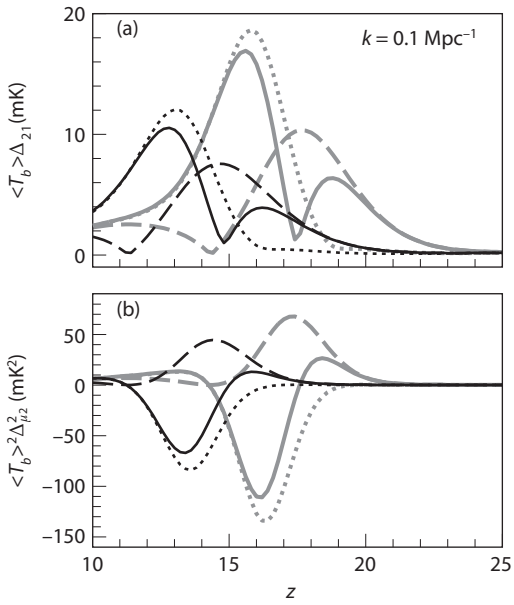


Figure 12.15 Evolution of brightness temperature fluctuations at $k = 0.1\,\mathrm{Mpc}^{-1}$ in two models of the early history of the 21-cm background; this is near the peak sensitivity of most of the planned experiments. The thick curves use parameters similar to our fiducial model in Figure 12.9 (and also the thick curves in Figure 12.8); the thin curves use the Population III model from Figure 12.8. The dashed curve includes the effects of Lyman-α fluctuations *only*, the dotted curve includes the effects of heating fluctuations *only*; the solid curve includes both (but *not* ionization). Panels *(a)* and *(b)* show the spherically averaged rms fluctuation and one of the anisotropic components, respectively. Pritchard, J. R., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **376**, 1680 (2007). Copyright 2007 by the Royal Astronomical Society.

light / AA

dark / AA

light / AA

snapshots of the fields (in the left column) and the corresponding (spherically averaged) power spectra (in the right column). The underlying model is very similar to the fiducial model whose mean signal is shown in Figure 12.9, though the redshifts of the critical points differ slightly. Importantly, the fluctuations are substantial throughout all the interesting regimes.

The top row of Figure 12.14 shows the point where Lyman-α pumping begins to be significant. The hydrogen gas is cold ($T_K \ll T_\gamma$), and the spin temperature is just beginning to decouple from the CMB. In this case the fluctuations are driven by the discrete, clustered first galaxies: their radiation field drives $T_S \to T_K$ around those first sources while leaving most of the IGM transparent.

In this model, the Lyman-α radiation field very quickly builds up the brightness temperature fluctuations. We illustrate this feature in Figure 12.15, which shows the evolution of the amplitude of the power spectrum at one particular

wavenumber ($k = 0.1$ Mpc^{-1}, near the peak sensitivities of most arrays).[xi] The dashed curve shows the effects of the Lyman-α fluctuations: they build up to a peak, with amplitude $\sim$10 mK, before decreasing again once Lyman-α fluctuations become strong everywhere (so that $\beta_\alpha \propto 1/x_\alpha \to 0$).

The second row in Figure 12.14 shows the signal near the onset of X-ray heating. At this point in the model, the Lyman-α coupling is strong nearly everywhere, so most of the IGM is cold and hence appears in absorption. But near the first X-ray sources (assumed to be star-forming galaxies here), the X-ray background has already heated the gas to $T_S \gg T_\gamma$, so these regions appear in emission. The net effect is a very large fluctuation amplitude, with a strong contrast between emitting and absorbing regions.

Figure 12.15 also helps to illustrate this behavior. Here, the dotted curve includes *only* the effects of heating fluctuations (implicitly assuming strong Lyman-α coupling throughout). The signal rises to $\sim$20 mK when this strong contrast is in place, then the fluctuations decrease once more of the IGM becomes hot (and hence saturates in emission).

as / AA

The solid curve in Figure 12.15 includes both heating and Lyman-α fluctuations. In this model the X-ray background lags the Wouthuysen-Field coupling, but not by a large margin. As a result, the net signal actually *decreases* in the early phases of the heating era. This occurs because only the regions near the first sources have strong coupling, but these are also the regions that are heated; the resulting emission signal is weaker than absorption because of the saturation in equation (12.8). Once the Lyman-α background reaches more of the IGM, the signal increases quickly.

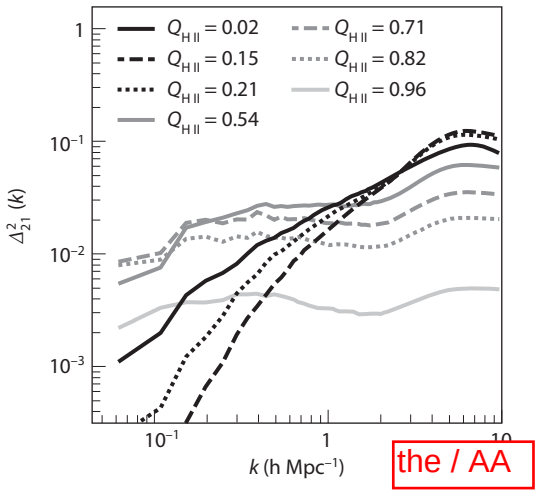
The third row in Figure 12.14 shows the 21-cm signal after heating has saturated ($T_S \gg T_\gamma$) throughout the IGM. At this point, spin temperature fluctuations no longer contribute to T_b, and only the density field affects the overall signal. The fluctuations are thus relatively modest (as in the late stages of the model of Figure 12.15). However, this period could be very important for cosmological measurements, because the astrophysical uncertainties in the ionized fraction and T_S are less significant (see later discussion).

Finally, the fluctuations increase again once reionization begins in earnest, as shown in the bottom row of Figure 12.14: here the fluctuations in the map are dominated by the contrast between the ionized bubbles and fully neutral gas between them. As we saw in §9.4, the pattern of these bubbles contains information about the ionizing sources creating them.

Figure 12.16 shows how the dimensionless power spectrum $\Delta^2_{21}(k) = k^3 P_{21}(k)/(2\pi^2)$ (or the power per logarithmic interval in wavenumber of the 21-cm signal) evolves in a radiative transfer simulation of the reionization process. (To recover the 21-cm signal one needs to multiply these values by the mean brightness temperature in a fully neutral medium, $T_0^2 \approx [28^2(1+z)/10]$ mK^2.) The different curves show a sequence of ionized fractions, from nearly neutral

[xi] This example is taken from a different analytic model, so the times at which the critical points differ relative to the seminumerical calculation; however, the qualitative evolution is identical.

of / AA



the / AA

Figure 12.16 Dimensionless power spectra $\Delta^2_{21}(k)$ of spin-flip background during the reionization era in a numerical simulation with radiative transfer; to obtain the 21-cm signal one needs to multiply $\Delta^2_{21}(k)$ by the mean brightness temperature in a fully neutral medium, $\sim[28^2(1+z)/10]$ mK^2. The curves show the power spectrum through a sequence of mean ionized fractions; the redshifts at which these points are achieved (not listed) do not significantly affect the signal, except through the mean brightness temperature. Lidz, A., et al., *Astrophys. J.* **680**, 982 (2008). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

($Q_{\rm H\,II} = 0.02$) to almost fully ionized ($Q_{\rm H\,II} = 0.96$). In this model, these span the range of redshifts $z \sim 11.5$–6.8, but the curves change little if one holds $Q_{\rm H\,II}$ constant but chooses a different redshift.

Clearly, both the shape and amplitude of the power spectrum evolve substantially throughout reionization. At first, the 21-cm power spectrum simply traces the matter power spectrum, as ionized regions have not yet significantly affected the IGM (and in this model $T_S \gg T_\gamma$ throughout the IGM, so spin temperature fluctuations are likewise unimportant). The power then *decreases* on large scales because the ionized bubbles appear first in the densest regions, suppressing the signal there and hence decreasing the overall contrast in the 21-cm maps.

This behavior is simplest to understand if we decompose the power spectrum into parts that describe perturbations in each relevant physical parameter and retain only the dominant components (see equation 12.42):

$$\Delta^2_{21}(k) = T_0^2\, Q^2_{\rm H\,I} \left[\Delta^2_{\delta\delta}(k) + 2\Delta^2_{x\delta}(k) + \Delta^2_{xx}(k)\right]. \qquad (12.54)$$

In this equation, $\Delta^2_{\delta\delta}$ and Δ^2_{xx} represent the power spectra of the density field and ionized fraction, and $\Delta^2_{x\delta}$ is the cross–power spectrum of the density with

the ionized fraction.[xii] Because $\Delta^2_{x\delta}$ is a cross power, it can be negative—that is, the neutral fraction $Q_{\rm H\,I}$ is small when δ is large in most reionization models. In the early phases of reionization, this term dominates the ionized power itself, Δ^2_{xx}, and so the net power falls.

However, by $Q_{\rm H\,II} \sim 0.5$, the ∼20 mK contrast between ionized and neutral gas dominates the maps, and the power increases rapidly: now the ionized bubbles fill a wide range of density, so $\Delta^2_{x\delta}$ is small but Δ^2_{xx} is large—at least on large scales. In fact, the power from this term peaks on the characteristic scale of the ionized bubbles (which is well defined in most reionization models; see Figure 9.5). In combination with the contribution from the matter power spectrum, this power peak leads to a strong enhancement of power on moderate scales ($k \sim 0.1\,{\rm Mpc}^{-1}$), followed by a decline at smaller wavenumbers (not shown clearly in this figure because of the finite size of the simulation box).

At the same time, on scales much smaller than the bubble size, the 21-cm power is significantly smaller than expected from the matter power spectrum alone. This difference is largely due to the higher-order terms we have ignored: within an ionized region, the ionized fraction is largely uncorrelated with the small-scale density perturbations. Effectively, then, the contrast on these scales decreases because many of the small-scale overdensities no longer appear in the 21-cm map. The net effect is an overall *flattening* in Δ^2_{21} throughout reionization. The flattening shifts to larger scales throughout reionization, and the amplitude decreases as less of the gas can emit 21-cm photons.

Because the power spectrum of the ionized fraction dominates the signal on large scales, the spin-flip background could be an effective tool for studying the morphology of reionization (and the sources that drive it); the shape and amplitude of the power spectrum can inform us of the time history of reionization throughout the IGM and (through the bubble size distribution) the clustering properties of the sources that drive it. This interpretation is relatively model-independent (in contrast with galaxy surveys, whose implications for reionization are difficult to interpret owing to the many unknown properties of the observed galaxies).

The final phase in the evolution of the 21-cm background is the end of reionization, when the vast majority of the gas is ionized, and so the spin-flip signal declines dramatically. But it does not disappear: substantial reservoirs of neutral gas still exist inside self-shielded galaxy-sized objects—the "damped Lyman-α absorbers" we discussed earlier. Observations show that these systems typically have $T_S \gg T_\gamma$; in this limit the power spectrum is simply

$$\Delta^2_{21}(k) \approx T_0^2\, Q^2_{\rm H\,I}\, \Delta^2_{gg}(k), \tag{12.55}$$

where Δ^2_{gg} is the galaxy power spectrum (which can be computed easily with the halo model), and $Q_{\rm H\,I}$ is measured from a census of DLAs to be a few percent

[xii] Here we have included only "low-order" terms in which two quantities are correlated, for simplicity. In fact, because the ionized fraction is usually either ∼0 or ∼1, "higher-order" terms such as $\Delta^2_{x\delta,x\delta}$, expressing the joint correlations between ionized fraction and density evaluated at two different locations, are not necessarily smaller than the terms we have retained.

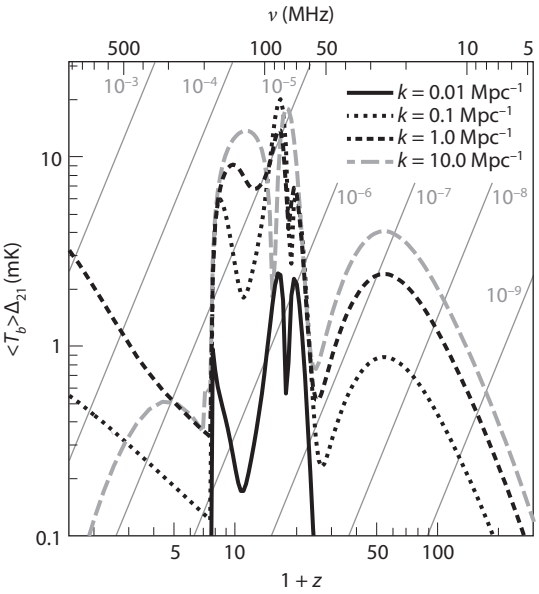


Figure 12.17 Redshift evolution of the angle-averaged 21-cm power spectrum in a model with reionization ending at $z = 6.5$. We show the amplitude for $k = 0.01\,\mathrm{Mpc}^{-1}$ (solid curve), $k = 0.1\,\mathrm{Mpc}^{-1}$ (dotted curve), $k = 1\,\mathrm{Mpc}^{-1}$ (short-dashed curve), and $k = 10\,\mathrm{Mpc}^{-1}$ (long-dashed curve). After reionization, the fluctuations trace neutral gas inside galaxies and DLAs and so mirror the galaxy power spectrum. The diagonal curves show contours of a fixed fraction of the sky brightness as a function of frequency. Pritchard, J. R., & Loeb, A., *Phys. Rev.* **D 78**, 103511 (2008). Copyright 2008 by the American Physical Society.

after reionization.[20] Figure 12.17 compares the postreionization signal with the higher-redshift one at several different wavenumbers in a model where reionization is tuned to end at $z = 6.5$. Provided galactic systems do dominate the neutral gas, fluctuations in the spin-flip background at redshifts after reionization therefore present an interesting cosmological probe—with the same information as galaxy surveys—but offer little information about the IGM itself.

12.6.1 Extracting Cosmological Measurements from the Spin-Flip Background

To this point, we have focused on the spin-flip background as a rich *astrophysical* data set. However, it also holds great promise for measurements of "fundamental" cosmological information, much like the CMB. There are several reasons for this promise. First, the 21-cm signal probes a time period when structure formation was still in its infancy—and, in particular, still within the well-understood linear regime through most of space. Second—unlike with galaxy surveys—the 21-cm signal probes the majority of baryonic matter that

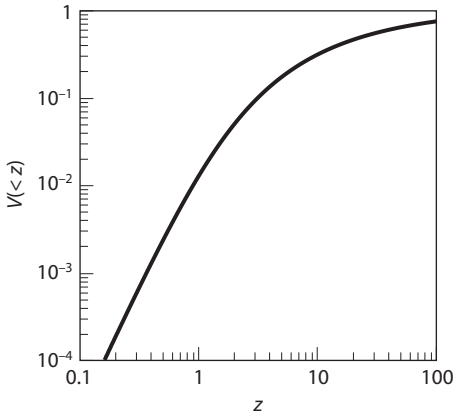


Figure 12.18 The fraction of the total comoving volume of the observable Universe that is available up to a redshift z, as a function of z. Loeb, A., & Wyithe, J.S.B., *Phys. Rev. Lett.* **100**, 161301 (2008). Copyright 2008 by the American Physical Society.

lies outside virialized structures, allowing us to access directly the linear fluctuations in the matter field.

Third—unlike the CMB—a 21-cm survey yields a three-dimensional dataset and hence probes a much larger fraction of the cosmic volume. We will see in the next section that the ultimate fractional uncertainty in the amplitude of any Fourier mode of wavelength λ is given by $\sim 1/\sqrt{N}$, where N is the number of independent elements of size λ that fit within the survey volume. For the two-dimensional map of the CMB, N is the surveyed area of the sky divided by the solid angle occupied by a patch of area λ^2 at $z \sim 10^3$. For a three-dimensional field, we obtain one of these maps at every frequency, vastly increasing the size of the available data set. Figure 12.18 shows the fraction of the total comoving volume of the observable Universe that is available up to different redshifts. Clearly, 21-cm surveys at $z \sim 10$ probe a much bigger comoving volume than conventional galaxy surveys at $z < 1$.

Finally, the 21-cm power extends down to the pressure-dominated (Jeans) scale of the cosmic gas. This is orders of magnitude smaller than the comoving scale at which the CMB anisotropies are damped by photon diffusion. Consequently, the spin-flip background can trace the primordial inhomogeneities with a much finer resolution (i.e., many more independent pixels) than the CMB. Altogether, the preceding factors imply that 21-cm tomography of cosmic hydrogen may potentially carry more information about the initial conditions of our Universe than any other method.

Of course, extracting cosmological information in the presence of the rich astrophysics that sets the 21-cm brightness may be challenging. Fortunately, there are two regimes in which it may be possible. The first is *before* the first stars lit up the Universe. During these dark ages, there was no astrophysics

that could possibly interfere; however, at such high red-shifts, corresponding to low frequencies, the noise is extraordinarily large (see Figure 12.17), so this era will remain inaccessible for the foreseeable future.

A second possibility is that the ionization and temperature factors in equation (12.8) vanish; then, the spin-flip brightness traces density and velocity fluctuations, both of which can easily be translated into fundamental cosmological parameters like the matter content or Hubble constant. We saw in §12.4 that such a scenario is plausible: the Wouthuysen-Field effect could have become strong long before reionization began in earnest, and X-ray heating could also have been very fast—but this outcome is by no means guaranteed. If not, cosmological information can be extracted only if the astrophysics is well understood.

We will see later that high-precision measurements of the 21-cm background are challenging, and for the foreseeable future the direct constraints on, for example, the matter power spectrum will not be competitive with those from galaxy surveys or the CMB. However, because the spin-flip background extends to such small scales, it still adds new cosmological information compared with other measurements. This tool is particularly useful for cosmological parameters that depend crucially on small scales, such as the shape of the primordial power spectrum and the neutrino mass (because the free streaming of neutrinos erases small-scale power).[21]

12.7 Mapping the Spin-Flip Background

The prospect of studying reionization, and even earlier epochs, 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 antennas. For redshifts $z \sim 6$–50, the corresponding observed frequencies are $\nu_{\rm obs} \sim 30$–$200\,{\rm MHz}$. Although the radio technology for the frequency range of interest has existed for decades—and is essentially the same that we use every day for television or radio communication—these experiments face three extreme challenges before they can observe the spin-flip background:

- The low-frequency band is heavily used by humans (as it includes the FM radio band, analog TV stations, and a host of satellite and aircraft communications channels), and the resulting *terrestrial radio interference* is as many as 10 orders of magnitude brighter than the 21-cm background. Most of the designs therefore place the observatories in isolated locations far from the contaminating sources (although some residual contamination does remain). However, this interference is usually (though not always) narrowband, so one can also attempt to measure the cosmological signal only in the gaps between contaminated channels. Even then, the presence of such bright foregrounds places serious requirements on the dynamic range of the low-frequency observatories.

- The *ionosphere* is refractive at low frequencies (and at the lowest frequencies, corresponding to redshifts $z > 50$, becomes opaque). This causes sources to jitter across the sky as patches of the ionosphere move across the telescope beam. The refraction phenomenon is similar to atmospheric seeing in optical astronomy, although the timescale for the jitter is much slower (several seconds in this case). It can be corrected in software by calibrating to the locations of a set of point sources distributed across the field of view, although this is by no means a trivial computing effort. The ionosphere is more active during the day and during times of high solar activity. This activity—together with the large brightness of the sun itself at these frequencies—restricts these observatories to operating only at night.
- Most significantly, the spin-flip background is far from the only astronomical source in the sky. Nearly all nonthermal radio sources are bright in the low-frequency band, especially the synchrotron radiation from the Milky Way galaxy, as we saw in Figure 12.11. But other extragalactic sources—including AGN, galaxy clusters, and even normal star-forming galaxies—also contribute. As rule of thumb, typical high-latitude, "quiet" portions of the sky have a brightness temperature[22]

$$T_{\rm sky} \approx 180 \left(\frac{\nu}{180\,{\rm MHz}}\right)^{-2.6} {\rm K}. \tag{12.56}$$

  This brightness is so large that it swamps the noise from even a simple receiver. We immediately see that 21-cm mapping will require large integration times and large collecting area to overcome this "noise," which is at least 10^4 times stronger than the cosmic signal.

Despite numerous efforts over the past four decades to observe fluctuations in the spin-flip background, these factors—as well as not-yet mature theories of the first galaxies—have conspired to prevent any detection. Now, with modern computing, it has become possible to analyze the enormous volume of data generated by experiments to see this background. Currently, several experiments are either in the early phases of operations or final phases of construction. All these mapping experiments are *interferometers*, in which the signals from multiple antennas are correlated to produce one larger, higher-resolution telescope.

The wide ranges of approaches taken by teams highlight the vitality of this field; the theoretical promise described in this chapter is now being transformed into actual instruments. The current tomographic projects include the following:

a decade /AA

- The Giant Metrewave Radio Telescope (GMRT; in India) is an interferometer with thirty 45-m antennas operating at low radio frequencies. The 21-cm background was an early motivator for the project, which was completed more than ~~10 years~~ ago, but the theoretical landscape changed radically, and only now has GMRT returned to this project. The large collecting area provides a powerful tool, but the instrument's narrow field of view and difficult radio environment present challenges. Nevertheless, the GMRT

team was the first to put limits on the spin-flip background, ruling out a cold, neutral IGM at $z \sim 8$ in the summer of 2010.[23]

- The Low Frequency Array (LOFAR; with the core in the Netherlands and outlying stations throughout Europe) is a large, general-purpose low-frequency radio telescope that began ~~science~~ operations in 2010. Because it has many other science goals, LOFAR is not completely optimized to observe the spin-flip background, but its large collecting area (especially inside a compact "core" most useful for these observations) and advanced computers nevertheless make it a powerful machine for this purpose. Its location in Western Europe means that LOFAR faces by far the most difficult terrestrial radio environment. Moreover, it uses an enormous number of dipole antennas and combines their individual signals into "stations" that are then used as interferometers. While this design allows for a large collecting area, it presents analysis challenges in understanding the instruments sufficiently well to extract the tiny cosmological signal.
- The Murchison Widefield Array (MWA) in Western Australia is an interferometer built almost entirely to observe the 21-cm background. Thus, the project hopes to leverage the relatively small experiment into limits competitive with larger first-generation experiments. Like LOFAR, MWA uses thousands of dipoles grouped into "tiles," which increase the collecting area at the cost of complexity. Because MWA's tiles are smaller, though, it achieves a larger field of view than LOFAR, which partially compensates for the much smaller collecting area. Figure 12.19 illustrates the antenna tile design of MWA.
- The Precision Array to Probe the Epoch of Reionization (PAPER, with instruments in Green Bank, West Virginia; and South Africa) combines signals from single dipoles into an interferometer. Without tiles, PAPER has a much smaller total collecting area than the other projects but has the advantages of a well-calibrated and well-understood instrument and an enormous field of view. The PAPER instrument is gradually building toward 128 antennas.

AA

In addition to this impressive suite of ongoing efforts, larger experiments are planned for the future, with their designs and strategies informed by this present generation.

In this section we briefly describe how these experiments work and hope to measure the spin-flip background. Of course, we cannot hope to do full justice to a topic as rich as radio observations and interferometry in this chapter, so we focus on the ideas most relevant to the spin-flip background and refer the interested reader to one of the many good textbooks on radio astronomy for more detailed information (see Further Reading).

DES: 2 lines under head ok?

12.7.1 A Brief Introduction to Radio Telescopes

The sensitivity of a telescope system depends on the competition between the strength of the cosmic signal collected by the antenna and the noise. The signal

Figure 12.19 One of the antenna "tiles" used in the Murchison Widefield Array (MWA) experiment in Western Australia (see *Color Plate 30* for a color version of this figure). Each such tile is composed of 16 crossed-dipole antennae, with their signals combined through hardware at the station. The full telescope combines the signals from ∼128 tiles interferometrically. This allows for a large (several hundred square degree) field of view with a moderately large collecting area. The antennas operate in the range 80–300 MHz, corresponding to $z \approx 6$–15 (although the telescope will be sensitive to the spin-flip background only at $z < 10$). Courtesy of / C. Lonsdale.

output of the antenna can be specified as an *antenna temperature*, T_a, which is the temperature of a matched resistive load that would produce the same power level ($P_a = k_B T_a \Delta\nu$ for the resistor) as the signal power $P = A_e S_\nu \Delta\nu/2$ received in one of two orthogonal antenna polarizations, where A_e is the effective collecting area of the telescope, S_ν is the source flux density (assuming an unpolarized source), and $\Delta\nu$ is the observed frequency bandwidth. From these terms we define the antenna sensitivity factor $K_a \equiv T_a/S_\nu = A_e/2k_B$.

The signal-to-noise ratio is evaluated by comparing T_a and $T_{\rm sys}$, the *system temperature*, which is similarly defined as the temperature of a matched resistor input to an ideal noise-free receiver that produces the same noise power level as measured at the output of the actual receiver. The system temperature includes contributions from the telescope, the receiver system, and the sky; the latter dominates in our case. Noise fluctuations ΔT^N decline with increased bandwidth and integration time $t_{\rm int}$ according to the radiometer equation,

$$\Delta T^N = \kappa_c \frac{T_{\rm sys}}{\sqrt{\Delta\nu\, t_{\rm int}}} \approx \frac{T_{\rm sys}}{\sqrt{\Delta\nu\, t_{\rm int}}}, \tag{12.57}$$

where $\kappa_c \geq 1$ is an efficiency factor accounting for the details of the signal detection scheme; for simplicity we set $\kappa_c = 1$, which is a reasonable approximation

for the telescopes discussed here. Equation (12.57) has a simple interpretation. Since the occupation number of the photons is large, they behave as a classical electromagnetic wave. The number of independent samples of the noise temperature is then the number of cycles observed during the integration time, $N_{\rm cyc} \sim \Delta\nu t_{\rm int}$, and the uncertainty in the system temperature $T_{\rm sys}$ is reduced by the factor of $\sqrt{N_{\rm cyc}}$ for Gaussian statistics (applicable in the limit of $N_{\rm cyc} \gg 1$).

The noise level (in flux density units) for an unresolved source is then

$$\sigma_S = \frac{T_{\rm sys}/K_a}{\sqrt{\Delta\nu\, t_{\rm int}}}. \tag{12.58}$$

Note that the noise level decreases with the telescope collecting area A_e. However, in many applications, we must take into account that the total collecting area may be distributed over a much larger physical area, to achieve better angular resolution $\theta_D \approx \lambda/D_{\rm max}$, where λ is the (observed) wavelength, and $D_{\rm max}$ is the maximum distance between antennas. In this case, the equivalent brightness temperature uncertainty is

$$\Delta T^N = \frac{\sigma_S c^2}{2k_B \nu^2 \Omega_B} \equiv \frac{T_{\rm sys}}{\eta_f \sqrt{\Delta\nu\, t_{\rm int}}}, \tag{12.59}$$

where $\eta_f \equiv A_{\rm tot}/D^2_{\rm max}$ is the *array filling factor*, and $\Omega_B \approx \theta_D^2$ is the solid angle subtended by the telescope beam. An appreciation of this dependence on η_f is crucial: the integration time required to detect a given surface brightness grows as $t_{\rm int} \propto D^4_{\rm max}$ if the (fixed) total collecting area is spread over larger areas to achieve better angular resolution.

We can develop better insight into the radio telescope response through a thought experiment in which a radio telescope is encased in a blackbody of temperature T. Regardless of its size, and with proper impedance matching, the telescope will produce an antenna temperature $T_a = T$ at its output. For this reason, attempts to observe the global 21-cm background are more concerned with issues of matching and gain calibration than with antenna size.

However, a telescope constructed with a beam of solid angle Ω_B will still deliver $T_a = T$ at its output if (*i*) it is embedded in a blackbody radiation field *or* (*ii*) an emitter of brightness temperature $T_B = T$ entirely fills its beam. Unfortunately, real radio telescopes do not form perfectly defined beams, and all suffer from sidelobes whose shapes and responses are dictated by diffraction and scattering of the incident radiation through the telescope. This is especially true of arrays, where a fraction $(1 - \eta_f)$ of the total response lies outside the beam defined by $\theta_{\rm D} \sim \lambda/D_{\rm max}$.

Using equation (12.56) with $T_{\rm sys} \approx T_{\rm sky}$ to estimate the telescope noise ΔT^N for a single-dish measurement of an unresolved source, we find

$$\Delta T^N|_{\rm sd} \approx 0.6{\rm mK} \left(\frac{1+z}{10}\right)^{2.6} \left(\frac{\rm MHz}{\Delta\nu}\frac{100\ {\rm hr}}{t_{\rm int}}\right)^{1/2}. \tag{12.60}$$

The mean 21-cm signal has $T_0 \sim 20\,{\rm mK}$; thus, single-dish telescopes can easily reach the sensitivity necessary to detect the global 21-cm background. In this

should be subscript italics f / AA

insert space / PE

regime, the challenge is instead to separate the slowly varying cosmological signal from the foregrounds. However, detecting individual features is still limited by the resolution of the telescope: a small single dish can detect the mean signal across the entire sky but cannot identify individual ionized bubbles.

12.7.2 Noise Estimates for 21-cm Interferometers

At radio frequencies, interferometry is required to make maps with even a relatively coarse resolution; for realistic collecting areas, the array dilution factor η_f dramatically decreases the sensitivity. Again using equation (12.56) for the system temperature, we find

$$\Delta T^N|_{\rm int} \sim 2\,{\rm mK} \left(\frac{A_{\rm tot}}{10^5\ {\rm m}^2}\right) \left(\frac{10'}{\theta_D}\right)^2 \left(\frac{1+z}{10}\right)^{4.6} \left(\frac{\rm MHz}{\Delta\nu}\,\frac{100\ {\rm hr}}{t_{\rm int}}\right)^{1/2}. \quad (12.61)$$

The angular resolution scale of $\theta_D \sim 10'$ and the frequency resolution scale of $\Delta\nu \sim 1$ MHz correspond to $\sim$20 Mpc.[xiii] The current generation of telescopes have $A_{\rm tot} < 10^5$ m^2, so imaging (i.e., mapping pixels with a signal-to-noise ratio much greater than unity) will be possible only on large scales that exceed the typical sizes of bubbles during most of reionization. It is for this reason that near-term imaging experiments focus primarily on giant H II regions generated by extremely luminous quasars during the middle phases of reionization, when the contrast between the large ionized bubble and the background IGM is largest.

Although equation (12.61) provides a simple estimate of an interferometer's sensitivity, we will see that the rate at which interferometers sample different scales depends on its design; this effectively makes η_f a function of angular scale. Thus, equation (12.61) provides only a rough guide.

When two antennas are coupled electronically to form an interferometer, the combined response projected on the sky resembles the characteristic diffraction pattern of a double slit. The spacing depends on the distance between the two elements, or the *baseline*. In general, the interferometer response to the sky brightness distribution $I_\nu(\hat{\mathbf{n}})$ for a particular "visibility" $\mathbf{V}$, corresponding to a particular baseline and frequency pair, in units of temperature, is

$$\mathbf{V}(\hat{\mathbf{n}}_0, u, v, \nu) \approx \int dx\, dy\, T_b(x, y, \nu) W_\nu(\hat{\mathbf{n}}_0, \hat{\mathbf{n}})\, e^{2\pi i\,(ux+vy)}, \quad (12.62)$$

where W_ν is the normalized response pattern of the antennas, and $\mathbf{A} = \lambda(u\hat{\mathbf{i}}, v\hat{\mathbf{j}}, w\hat{\mathbf{z}})$ is the vector (on the ground) between the two elements. In the orthogonal (u, v, w) coordinate system, the w-axis aligns with the direction toward the sky at the center of the beam, and the u- and v-axes are oriented so that the v-axis projects onto the local meridian. The coordinates x and y are angles measured in the "sky plane" relative to the intersection of $\hat{\mathbf{z}}$ with the celestial sphere. In this Fourier transform of the sky, u and v represent spatial frequencies, and

[xiii] More precisely, a bandwidth $\Delta\nu$ corresponds to a comoving distance $\sim$1.8 Mpc$(\Delta\nu/0.1\,{\rm MHz})[(1+z)/10]^{1/2}$, while an angular scale θ_D corresponds to $2.7(\theta_D/1')[(1+z)/10]^{0.2}$ Mpc.

the w-axis produces a phase offset in the interferometer fringe that can be calibrated. (This representation assumes that the interferometer sees only a small piece of the sky, so that the "flat sky" approximation is valid; that is not true for some of the 21-cm telescopes, but the basic formalism presented here provides a reasonable approximation with much less technical difficulty.)

We must keep in mind that this Fourier integral does not properly account for sources far outside the primary beam; in effect, these add a noise-like contribution entering through the sidelobes that inevitably appear outside the primary beam.

Given the difficulty of high signal-to-noise imaging, attention has focused on statistical measurements. We now turn to estimating the sensitivity of 21-cm experiments to the power spectrum. Error estimates for other statistical measures must still be developed, but the basic principles are the same. For simplicity, we consider only the effects of thermal noise and cosmic variance, which provide a fundamental limit. Systematics (especially foregrounds) present equally large difficulties, and the community is hard at work developing strategies to mitigate them, some of which we will discuss later.

We begin with the complex visibility of equation (12.62). The detector noise for a single visibility measurement is closely related to equation (12.57). Equation (12.58) implies

$$\Delta T^N(v) = \frac{\lambda^2\, T_{\rm sys}}{A_e\,\sqrt{\Delta v t_{\mathbf{u}}}}, \tag{12.63}$$

where here, $t_{\mathbf{u}}$ is the integration time of this particular baseline; owing to Earth's rotation, these large interferometers continually shift their sky coverage (in a manner analogous to "drift-scanning" in optical astronomy), so this is *not* the same as the total integration time. Also, A_e is the collecting area of each antenna element (which we assume to be perfectly efficient, for simplicity).

The observed "visibility data cube" conventionally used in radio astronomy is actually a hybrid of Fourier-space (u, v) and redshift-space (v) coordinates and is thus inconvenient for comparing with theoretical models. One can either transform the visibility data to the sky plane to obtain the "image cube" or transform the frequency (redshift) coordinate to its Fourier-space equivalent[xiv] to obtain a representation with spatial frequency for all three dimensions,

$$T_b(\mathbf{u}) = \int_B dv\, \mathbf{V}(u, v, v)\, e^{2\pi i \eta v}, \tag{12.64}$$

where the integration extends over the full bandwidth B of the observation, $\mathbf{u} \equiv u\hat{\mathbf{i}} + v\hat{\mathbf{j}} + \eta\hat{\mathbf{z}}$, and η has dimensions of time. In this representation, the effective noise can be obtained by Fourier transforming the signal across the frequency axis, which yields

$$\Delta T^N(\mathbf{u}) = \frac{\lambda^2\, T_{\rm sys}\,\sqrt{B}}{A_e\,\sqrt{t_{\mathbf{u}}}} \approx \frac{T_{\rm sys}}{\sqrt{B\, t_{\mathbf{u}}}} \times \frac{\lambda^2}{A_e\,\delta\eta}. \tag{12.65}$$

[xiv] Obviously, this transformation is not useful for noncosmological radio astronomy applications, where there is no frequency–distance relation.

In the second equality, we have set $\delta\eta = B^{-1}$. The factor $A_e/\lambda^2 \times \delta\eta$ then represents the Fourier-space resolution of the observation (or the inverse volume sampled by the primary beam, in the appropriate units); note the similarity to equation (12.57). Here $\Delta T^N(\mathbf{u})$ has units of temperature divided by time, because of the Fourier transform in the frequency direction.

To estimate the statistical errors, we need the covariance matrix of the noise for antenna pairs at baselines $\mathbf{u}_i$ and $\mathbf{u}_j$. Because the thermal noise errors are uncorrelated between measurements, this is simply a diagonal matrix in which each element is the square of equation (12.65). In transforming to the physical wavevector $\mathbf{k}$, we distinguish between the component $\mathbf{u}_\perp$ oriented along the sky (corresponding to $\mathbf{k}_\perp = 2\pi\mathbf{u}_\perp/D$, where D is the comoving distance to the observed survey volume) and the component $\mathbf{k}_\parallel$ along the line of sight. It is useful to do this because interferometers can have arbitrarily good frequency resolution, while the $\mathbf{u}_\perp$ coverage is always fixed by the baseline distribution.

We define the number density of baselines that observe a given $\mathbf{u}_\perp$ as $n(\mathbf{u}_\perp)$; this value is normalized so that its integral over the $u_\perp$ half-plane is $N_B = N_a\,(N_a - 1)/2$, the total number of baselines in the array of N_a antennas. Two properties of $n(\mathbf{u}_\perp)$ are noteworthy. First, because of Earth's rotation, it is azimuthally symmetric and thus only a function of $u_\perp = |\mathbf{u}_\perp|$. Second, for a smooth antenna distribution, $n(u_\perp)$ is virtually always a decreasing function of $u_\perp$. This fact follows from a simple geometric consideration: it is difficult to arrange the antenna distribution to have many more long baselines than short ones. We can write

$$t_{\mathbf{k}} \approx n(u_\perp) \left(\frac{A_e}{\lambda^2}\right) t_{\rm int}. \tag{12.66}$$

As before, $A_e/\lambda^2 \approx \delta u\,\delta v$ is the angular component of the Fourier-space resolution. Thus, the noise covariance matrix is

$$\begin{aligned} C^N(\mathbf{k}_i,\,\mathbf{k}_j) &\equiv \langle \Delta T^N(\mathbf{u}_i)^*\;\Delta T^N(\mathbf{u}_j)\rangle \\ &= \left(\frac{\lambda^2\,B\,T_{\rm sys}}{A_e}\right)^2 \frac{\delta_{ij}}{B\,t_{\mathbf{k}}}. \end{aligned} \tag{12.67}$$

Equation (12.67) represents the thermal noise contribution to the covariance matrix; even in an ideal experiment with no systematics from foregrounds, we must also include errors from sample variance. This component is

$$\begin{aligned} C^{SV}(\mathbf{k}_i,\,\mathbf{k}_j) &= \langle T_b^*(\mathbf{k}_i)\,T_b(\mathbf{k}_j)\rangle \\ &\approx \delta_{ij} T_0^2 \langle x_{\rm HI}{}^2\rangle \int d^3\mathbf{u}\,|\tilde{W}(\mathbf{u}_i - \mathbf{u})|^2\,P_{21}(\mathbf{u}) \\ &\approx T_0^2 \langle x_{\rm HI}{}^2\,P_{21}(\mathbf{k}_i)\rangle \frac{\lambda^2\,B^2}{A_e\,D^2\,\Delta D}\,\delta_{ij}, \end{aligned} \tag{12.68}$$

where $\Delta D \propto B$ is the line-of-sight depth of the observed volume in comoving units, and T_0 is the average brightness temperature of a fully neutral IGM. In the first line of the equation, the average is over baseline and frequency pairs indexed by $\mathbf{k}_i$ and $\mathbf{k}_j$ (or equivalently $\mathbf{u}_i$ and $\mathbf{u}_j$). In the second line, $\tilde{W}$ is the

Angle bracket goes before superscript 2 / PE

insert right angle bracket before superscript 2 / delete right angle bracket after (ki) / PE

Fourier transform of the primary beam response function, including the finite bandwidth, and is most naturally expressed in the "observed" units **u**. It typically differs from zero in an area $\delta u\,\delta v\,\delta\eta \approx A_e/(\lambda^2 B)$ and (ignoring efficiencies) integrates to unity over the beam. For the last line, we have assumed that **u** is much larger than the width of this response function. Then, $P_{21}(\mathbf{u})$ is constant across the beam and can be extracted from the integral, which becomes simply $(\delta u\,\delta v\,\delta\eta)^{-1}$. We have also transformed to the more physically relevant wavenumber **k**, which introduces a factor $B/(D^2\Delta D)$.

Angle bracket goes before superscript 2 / PE

Equation (12.68) has a simple physical interpretation: it is essentially a normalization factor $(T_0^2\langle x_{\rm H\,I}^2\rangle B^2)$ multiplied by $P_{21}/V_{\rm surv}$, where $V_{\rm surv} \approx \Delta D\,(D^2\lambda^2/A_e)$ is the total volume observed by the telescope. This factor counts the number of independent estimates N available to the measurement of a given Fourier mode; the squared error then scales as $1/N$.

delete comma, insert period, capitalize W / PE

To translate these values into error estimates, we use the common Fisher information matrix approach (see also §10.4.3), which provides an idealized estimate of the measurement errors given the total covariance matrix $\mathbf{C} = \mathbf{C}^N + \mathbf{C}^{SV}$; we ignore any possible systematics and inefficiencies in the data reduction. Given a vector of parameters Ψ, the $(i,\ j)$ element of the Fisher matrix is defined as[24]

$$F_{ij} \equiv \left\langle -\frac{\partial^2 \ln\mathcal{L}}{\partial\Psi_i\,\partial\Psi_j}\right\rangle \tag{12.69}$$

$$= \mathrm{Tr}\left[\mathbf{C}^{-1}\,\frac{\partial\mathbf{C}}{\partial\Psi_i}\,\mathbf{C}^{-1}\,\frac{\partial\mathbf{C}}{\partial\Psi_j}\right], \tag{12.70}$$

where $\mathcal{L}$ is the log-likelihood function. For the simple case of measuring the binned power spectrum from the data points, the "parameters" are the power spectrum amplitudes in each of the bins, $\Psi_i = P_{\Delta T} \equiv T_0^2\langle x_{\rm H\,I}^2\rangle P_{21}(\mathbf{k}_i)$; in more general cases they are the parameters of a theoretical model meant to describe the data. The *Cramer-Rao inequality* states that the errors on any unbiased estimator of the power spectrum must satisfy

insert right angle bracket before superscript 2 / PE

$$\delta P_{\Delta T}(\mathbf{k}_i) \geq \frac{1}{\sqrt{N_c(\mathbf{k}_i)}}\sqrt{(\mathbf{F}^{-1})_{ii}}, \tag{12.71}$$

where N_c is the number of measurements in the appropriate bin and $\mathbf{F}^{-1}$ is the inverse of the Fisher matrix.

In the case we are studying, the Fisher matrix is particularly simple to use because the covariance matrix is diagonal. (This will not be true for real data, because foreground cleaning and other systematic effects induce correlated residual errors, but the matrix provides a rough estimate of the noise limits.) The resulting error (from a single baseline) on a power spectrum estimate is

$$\delta P_{21}(\mathbf{k}_i) = P_{21}(\mathbf{k}_i) + \frac{T_{\rm sys}^2}{B t_{\rm int}}\,\frac{D^2\Delta D}{n(k_\perp)}\left(\frac{\lambda^2}{A_e}\right)^2. \tag{12.72}$$

The last step is to count the number of Fourier cells in each power spectrum bin, which depends on the Fourier-space resolution of the instrument. Recall

that when redshift-space distortions are included, P_{21} is not truly isotropic, but it is azimuthally symmetric. Thus, we use Fourier cells grouped into annuli of constant $(k,\ \mu)$. Then, in the limit that $k\mu$ is much larger than the effective $k_\parallel$ resolution,

$$N_c(k) \approx 2\pi\, k^2\, \Delta k\, \Delta\mu \times \left[\frac{V_{\rm surv}}{(2\pi)^3}\right], \tag{12.73}$$

where the last term represents the Fourier-space resolution. The total errors from all estimates within a bin simply add in quadrature. In its essence, this calculation is identical to our estimate for the errors on galaxy power spectrum measurements in §10.4.3, except that the shot noise relevant to galaxy number counts is replaced by the thermal noise of each mode.

Equations (12.67), (12.68), and (12.72) fully specify the effects of noise in the absence of systematic effects. But to make estimates we must determine the effective observing time $t_{\mathbf{k}}$ for each mode—and hence the baseline distribution $n(u_\perp)$ by equation (12.66)—as well as the sampling density (equation 12.73 for a measurement in annuli). These two quantities are obviously highly dependent on the design of the experiment. It is therefore useful to consider the simple thermal noise-dominated case to develop some intuition for array design. Substituting for N_c in equation (12.72) and ignoring the first term (which is equivalent to working on small scales), we find

$$\delta P_{\Delta T} \propto A_e^{-3/2}\, B^{-1/2} \left[\frac{1}{k^{3/2}\, n(k,\ \mu)}\right] \left(\frac{T_{\rm sys}^2}{t_{\rm int}}\right). \tag{12.74}$$

Here we have assumed that the power spectrum is measured in bins with constant logarithmic width in k but constant linear width in μ. From equation (12.74), we can deduce a number of fundamental considerations driving array design.

- First, $\delta P_{21} \propto t_{\rm int}^{-1}$, because the power spectrum depends on the square of the intensity.
- Second, we can increase the collecting area in two ways. One is to add antennas while holding the dish area A_e constant. Recall that $n(k,\ \mu)$ is normalized to the total number of baselines $N_B \propto N_a^2$: thus, adding antennas of a fixed size decreases the errors by the total collecting area squared. (Of course, the number of correlations needed also increases by the same factor, so this strategy is costly in terms of computing.) The other method is to make each antenna larger but hold their total number fixed. In this case, the total number of baselines, and hence $n(k,\ \mu)$, remains constant, but $\delta P_{\Delta T} \propto A_e^{-3/2}$. Increasing the collecting area in this way is not as efficient because it decreases the total field of view of the instrument, which is set by the field of view of each antenna.
- Third, adding bandwidth increases the sensitivity relatively slowly: $\delta P_{\Delta T} \propto B^{1/2}$, because it adds new volume along the line of sight without affecting the noise on any given measurement. Of course, one must be wary

of adding too much bandwidth because of systematics (especially foregrounds).

- Finally, as a function of scale k, $\delta P_{\Delta T} \propto k^{-3/2}\, n(k,\, \mu)^{-1}$. The first factor comes from the increasing (logarithmic) volume of each annulus as k increases. But in realistic circumstances the sensitivity actually decreases toward smaller scales because of n. This result is most obvious if we consider a map at a single frequency. In that case, high-k modes correspond to small angular separations or large baselines; for a fixed collecting area the array must therefore be more dilute, and the sensitivity per pixel decreases, as in equation (12.61). In the (simple but unrealistic) case of uniform uv coverage, the error on a measurement of the angular power spectrum increases as θ_D^{-2} for a fixed collecting area.

  Fortunately, the three-dimensional nature of the true 21-cm signal moderates this rapid decline toward smaller scales: even a single dish can measure structure along the line of sight on small physical scales. Mathematically, because $n(k,\, \mu) = n(k_\perp)$, each baseline can image arbitrarily large $k_\parallel$, at least in principle. For an interferometer, this implies that short baselines still contribute to measuring large-k modes. Thus, provided they have good frequency resolution, compact arrays are surprisingly effective at measuring small-scale power . There is one important caveat: if short wavelength modes are sampled only along the frequency axis, only modes with $\mu^2 \approx 1$ can be measured. Thus little, if any, information is recovered on the μ dependence of the redshift-space distortions. Studying this aspect of the signal *does* require baselines able to measure the short transverse modes with $\mu^2 \approx 0$.

Figure 12.20 summarizes the expected errors (including only thermal noise and cosmic variance, not systematics) on the spherically averaged power spectrum. The thin curves show a forecast for an experiment four times larger than the MWA (but otherwise with identical parameters, comparable to the best constraints expected from the first generation of experiments). We assume 1,000 hours of integration on a single field (roughly 1-year of realistic observing conditions), the observed modes ~~binned~~ into segments of width $\Delta k = k/2$ (as shown by the horizontal bars), and a radial survey width corresponding to 6 MHz (or $\Delta z \sim 0.5$).

binned / PE

Provided it reaches this limit, such an experiment can place fairly stringent constraints at scales of $k < 1\mathrm{Mpc}^{-1}$. Smaller scales are swamped by thermal noise. The errors on large scales come from cosmic variance, although here they are quite modest because of the large field of view of the telescope. To reach smaller scales will require more collecting area to reduce the noise. The thick curves show the estimated errors for a futuristic experiment, with 5,000 antenna tiles (40 times more than the MWA) and 400 times larger total collecting area. This experiment would provide good constraints out to $k \sim 10\mathrm{Mpc}^{-1}$.

insert space / PE

insert space / PE

Unfortunately, measuring very large physical scales with these experiments is likely to be very difficult, because of the other astronomical foregrounds. To separate the Galactic synchrotron radiation from the cosmological signal,

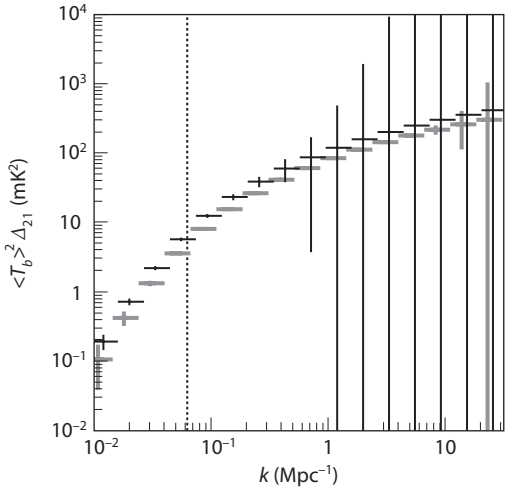


Figure 12.20 Estimated errors on the spherically averaged 21-cm power spectrum at $z = 8$ for an experiment four times larger than the MWA (thin curves) and one with 100 times larger collecting area (thick curves). The central values of the latter are shifted downward for clarity of presentation. We assume 1,000 hours of integration on a single field, the observed modes binnel into segments of width $\Delta k = k/2$ (the horizontal bars show these bins), and a radial survey width corresponding to 6 MHz (or $\Delta z \sim 0.5$). The vertical dotted line shows the scale corresponding to this bandwidth; foreground removal will likely prevent measurements at wavenumbers smaller than this scale.

the experiments will rely on the former's spectral smoothness and the latter's rapid variations with frequency (due to H II regions, density fluctuations, or temperature variations). The essential idea is to fit a low-order function to each pixel in the map (or Fourier mode) and subtract this mean variation over a wide (several megahertz) frequency range. Provided the foregrounds are smoother than the signal, this scheme will isolate the spin-flip background but with an inevitable loss of information (i.e., any variations in the 21-cm background over large frequency ranges will also be subtracted). Current estimates suggest that this method will work very well at small scales but will prevent measurements of any fluctuations on scales larger than those corresponding to the several-megahertz bandwidth of each measurement. The vertical dotted line in Figure 12.20 shows the scale corresponding to the assumed 6 MHz bandwidth; modes to the left of this line are likely lost in the foreground removal process. Unfortunately, this drastically reduces the dynamic range of the first-generation experiments.

Because the sky noise increases rapidly with redshift (see equation 12.56), the first generation of experiments lose sensitivity at $z \sim 11$–12. To reach these high redshifts will likely require collecting areas approaching a square kilometer. Such large instruments will also be necessary to measure the redshift-space

distortions in the spin-flip background, because they require separate measurements of power along the line of sight and across the plane of the sky. The first-generation experiments are relatively small and do not have adequate sensitivity to make high-resolution measurements on the plane of the sky, although they can do so in the redshift direction with reasonably narrow frequency channels. Much larger instruments are necessary to build sensitivity to fluctuations on the plane of the sky.

In addition to the unavoidable problems posed by foreground cleaning, there are several other serious systematic challenges to reaching the limits suggested by Figure 12.20. These include the ionospheric refraction described earlier, the many bright astronomical point sources (and especially their sidelobe contamination to the antenna beam), the variation of the instrument properties with frequency, and the polarized component of the foregrounds (which can vary rapidly with frequency and hence escape the typical foreground removal algorithms). Fortunately, several experiment teams are tackling these problems in unique ways, and the community hopes they can be overcome in the near future.

Chapter Thirteen

Other Probes of the First Galaxies

So far we have discussed four classes of observational probes of the first galaxies: direct observations of individual galaxies (over a variety of wavelengths), the Lyman-α line (both as a test of the galaxy populations and the IGM), the spin-flip line from intergalactic gas, and gravitational waves from black hole mergers (see §7.7). However, there are many other, less direct, ways to probe structures during the cosmic dawn. In this chapter, we discuss several of these:

- *Secondary anisotropies* of the CMB were generated as those photons passed through gas during the cosmic dawn. CMB photons did not interact with the IGM gas until it was ionized (with the exception of resonant lines); however, once that occurred, the photons began to scatter off the free electrons. The scattering process induced both large-scale polarization and small-scale temperature anisotropies.
- *Diffuse backgrounds* from the cosmic dawn (other than the spin-flip background) could have resulted from the integrated emission of the entire galaxy population. Typically, these backgrounds include galactic emission lines, ranging from CO lines in the radio to the Lyman-α line itself, so (like the spin-flip background) they contain not only angular structure but also spectral structure. Measuring these integrated backgrounds via low-resolution observations can be much easier than detecting individual galaxies (though of course also provides less information) and can quantify useful properties of the global galaxy populations.

  In the absence of detections from individual galaxies, the *cross-correlation* of different probes can help isolate cosmological information in the presence of contaminants and can often isolate interesting aspects of these diffuse signals.
- *Fossil structure* from early galaxies remains in (or can be deduced from) the Milky Way or other nearby entities in the Local Group. This fossil structure includes the residual effects of feedback (from the Lyman-Werner background, photoheating, or other processes) on the small satellite galaxies or globular clusters of the Milky Way, old low-mass stars that may have formed during the cosmic dawn and survive inside the Milky Way (or its halo), and remnant signatures of the early merger history of the Milky Way.

13.1 Secondary Cosmic Microwave Background Anisotropies from the Cosmic Dawn

DES: 2 lines under head ok?

The CMB indicates that hydrogen atoms formed 400,000 years after the Big Bang, as soon as cosmological expansion cooled the gas below 3,000 K. Once

neutral, the CMB photons could interact with the IGM gas only through its resonant transitions—and after $z \sim 1000$, when the photons redshift out of the Lyman-α resonance, the only available transition is the 21-cm line, whose effects we have already examined.

However, once the first stars or black holes began to ionize the IGM, the CMB photons scattered off the free electrons. This scattering had several effects on the CMB, which we describe in detail here.

13.1.1 Large-Scale Polarization of the CMB

The crucial parameter in determining these effects is the total CMB optical depth to electron scattering,

$$\tau_{\rm es} = \int n_e(z)\sigma_T (c\,dt/dz)\,dz, \tag{13.1}$$

where $\sigma_T = 6.65 \times 10^{-25}\,{\rm cm}^2$ is the Thomson scattering cross section, and the integral is over the path taken by a photon; note that only redshifts where the ionized fraction is nonzero will contribute (the residual ionized fraction following recombination produced only a very small contribution to $\tau_{\rm es}$, so the integrand is significant only once reionization began). In the simplest approximation, where we assume that the IGM was instantaneously ionized at $z_{\rm reion}$, equation (13.1) can be integrated analytically for a flat universe with $\Omega_\Lambda + \Omega_m = 1$:

$$\tau_{\rm es} = 4.44 \times 10^{-3} \times \left\{[\Omega_\Lambda + \Omega_m(1+z_{\rm reion})^3]^{1/2} - 1\right\}. \tag{13.2}$$

The coefficient here assumes for simplicity that helium was singly ionized at the same time as hydrogen; in reality, the second ionization of helium at $z \sim 3$ (see §4.5) adds an additional $\tau_{\rm es,He} \approx 1.02 \times 10^{-3}$.

The most obvious effect of this scattering on the CMB temperature anisotropies is to partially wash them out, as a fraction $e^{-\tau_{\rm es}}$ of the photons that appear (to the observer) to be incident from a particular direction actually come from elsewhere. Each line of sight samples only photons from a finite surrounding region, roughly a cylinder whose radius equals the causal horizon at the time of scattering. Thus, the angular power spectrum of the CMB fluctuations, C_ℓ—which contains two factors of temperature—is damped by a factor $e^{-2\tau_{\rm es}}$ on angular scales smaller than the causal horizon at ~~that~~ the time most scattering occurs, as shown in Figure 13.1.[1] (More precisely, electrons at each radius from the observer damp fluctuations across their local horizon, with the damping proportional to the scattering optical depth contributed by these electrons.) On larger scales, the power spectrum is unaffected. Unfortunately, this slight change to the slope of the temperature anisotropy power spectrum is strongly degenerate with the intrinsic tilt of the matter power spectrum.

the / AA

However, the scattering process also induces polarization, which is observable.[2] Consider photons that scatter off an electron toward the observer, as shown in Figure 13.2 (where the observer coincides with the reader, out of the page). Photons incident from the horizontal direction can scatter toward

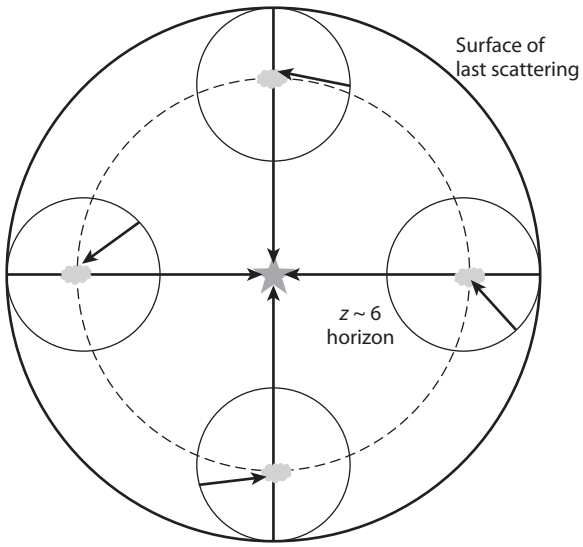


Figure 13.1 Diagram of the damping of primordial CMB anisotropies after reionization (not to scale). The observer is represented by the star at the center; the thick outer circle represents the surface of last scattering of the observer. CMB photons (arrows) propagate through a screen of parcels of free electrons at $z \sim 6$ (located at the dashed circle for the observer). Each parcel sees the CMB out to its causal horizon, represented by the smaller circle surrounding it. The low optical depth of free electrons implies that most CMB photons propagate through without interaction (long arrows), but some photons are scattered *into* the line of sight from other directions, which partially washes out the primordial anisotropies on scales smaller than the horizon at the time of scattering.

the observer only if they are polarized along the plane of the page in the vertical direction, because the other possible polarization state of the incident wave—out of the page—would not produce a scattered transverse wave. In contrast, to produce a transverse wave, photons incident from the vertical directions can scatter toward the observer only if they are polarized horizontally in the figure (along the plane of the page).

If we suppose that an electron scatters photons from all directions, it will produce a mixture of these horizontal and vertical polarization states. However, if there is an asymmetry in the incident radiation field—in particular, a quadrupole anisotropy between the intensity of the incident background along the horizontal and vertical axes—the resulting mixture will have a net polarization.

For the CMB, electrons during the cosmic dawn could scatter any photons incident on them that originated within the causal horizon at that time. The net polarization is $P \sim 0.1\tau_{es} Q^3$, where Q is the primordial quadrupole anisotropy on that horizon scale (the prefactor 0.1 comes from the detailed physics of the

move superscript 3 after is (so it does not get confused as the power of Q /AA

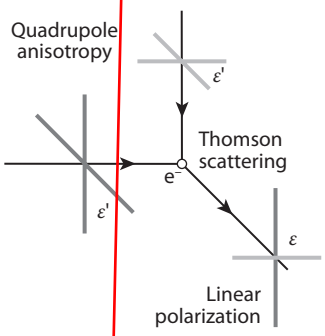


Figure 13.2 Cartoon of the generation of CMB polarization through Thomson scattering. The central electron scatters photons from the CMB. The scattering process allows only photons polarized perpendicular to the plane defined by the lines joining the observer and electron, and the photon's source and the electron. For photons from the left of the page, this is the vertical polarization, which corresponds to the thick vertical line at the lower right. For photons from the top of the page, this is the horizontal component (thin horizontal line at lower right). If the incident photon field has a quadrupole anisotropy (here, a larger intensity from the left), the resulting scattered field will be polarized. Hu, W., & White, M., *New Astron.* **2**, 323 (1997). Copyright 1997 by Elsevier.

Thomson scattering process). The overall amplitude of the polarization (or its power spectrum) therefore provides a measure of τ_{es} and hence the integrated column of ionized gas between us and the recombination surface.

The first such measurements have been made with the Wilkinson Microwave Anisotropy Probe (WMAP) over the past several years. The current estimate is $\tau_{es} = 0.087 \pm 0.017$, which would correspond to instantaneous reionization at $z \sim 10$.[4] This value will be refined by forthcoming CMB data from the Planck satellite.[5] Unfortunately, τ_{es} on its own does little to distinguish different reionization histories, as it simply measures the total probability of scattering between the present day and the surface of last scattering, without revealing how long ago this scattering occurred.

To learn about this history, we must turn to the scale dependence of the polarization or, equivalently, the shape of its power spectrum. The polarization anisotropies appear to us on the angular scale subtended by the horizon distance at the time of scattering, which occurs at $\ell < 40$. On much finer angular scales, the postreionization scattering washes out any primordial polarization, just as it does the temperature anisotropies. However, because there is no primordial polarization on large scales anyway, the "reionization bump" from the late-time scattering is very clearly distinguishable from primordial anisotropies.

Interestingly, because this horizon scale evolves with redshift, there is some information about the time history of $\bar{x}_i(z)$ contained in the CMB polarization power spectrum. Figure 13.3 shows some example reionization models (left) and their observable signatures (right). The models were generated with

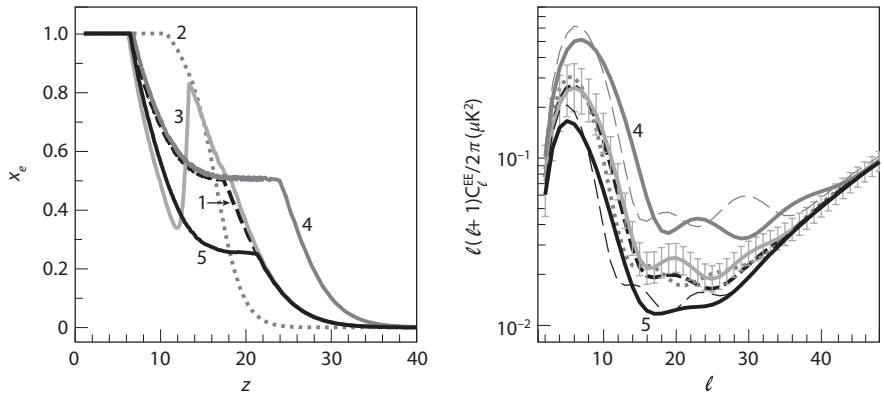


Figure 13.3 *Left:* Five example models of the reionization history. Models 1–3 are normalized to have $\tau_{es} = 0.17$, while models 4 and 5 have $\tau_{es} = 0.14$ and 0.23, respectively. Note that all these values are well above the current best estimate. See text for a discussion of the input physics in these models. *Right:* CMB polarization power spectra for these five models (shown is the EE component, which includes scalar perturbations). All are normalized to have the same power at $\ell > 50$, where reionization has no significant effect. The bold dashed, dashed, and solid lines correspond to models 1–3; the last has cosmic variance error bars attached. The light solid lines represent models 4–5, while the light dashed lines show best-fit polarization spectra for instantaneous reionization models. Holder, G., & Haiman, Z., *Astrophys. J.* **595**, 13 (2003). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

simple prescriptions for the physics of the first galaxies, but their particular physical assumptions are not important for our purposes; they simply provide a set of contrasting reionization histories to gauge its effects on the shape and amplitude of the polarization power.[i] Models 1–3 hold $\tau_{es} = 0.17$ but vary $\bar{x}_i(z)$. Models 4 and 5 show larger and smaller overall optical depths (of 0.23 and 0.14, respectively). Note that all these values are considerably larger than the current best estimate from WMAP but are shown here only to illustrate the dependence of the polarization signal on the reionization history.

The right panel shows the corresponding power spectra of the polarization anisotropies (in detail, it shows the *EE*, or scalar, component, which contains most of the contributions from reionization). Models 1–3 ~~(shown by the dark lines)~~ have similar amplitudes but different shapes, particularly around the trough at $\ell \sim 20$–30; the error bars, which give the ideal cosmic variance errors, show that these models can at least in principle be separated at high confidence. It is much easier to separate models with different optical depths: models 4 and 5 are shown by the ~~light~~ solid lines. Here we can also see the

AA

top and bottom / AA

[i]Model 3 in particular is physically implausible, because a nonmonotonic $\bar{x}_i(z)$ requires very strong, instantaneous feedback to suppress galaxy formation.

effect of $\bar{x}_i(z)$ on the power spectrum: the light dashed lines show the best-fit instantaneous reionization models, which provide relatively poor fits to the more complex reionization histories used in these models.

The reionization era also generates anisotropies in the B-mode polarization (so called because they have nonzero curl, as opposed to the E-modes, which are curl free). This type of anisotropy is of particular astrophysical interest because it is also generated by the gravitational wave background from the inflationary era, and there are numerous efforts underway to measure the background's properties. Secondary fluctuations, like those generated by reionization, are therefore important contaminants to understand. Fortunately, it appears that the B-modes generated by reionization are small and have large angular coherence in the polarization direction, which make them relatively easy to isolate.[6] Moreover, the 21-cm background can be used to reconstruct the electron-scattering optical depth along different lines of sight and from that information to reconstruct the expected polarization pattern (see §13.3).[7] Thus, B-mode anisotropies from reionization should not pose a substantial challenge to the detection of primordial gravitational waves.

The Planck satellite should be able to distinguish different reionization scenarios using the polarization anisotropies over the next several years. However, this technique is sensitive only to the global reionization history, not to the details of the reionization process. To understand the growth and morphology of the ionized regions, we must probe much smaller angular scales.

13.1.2 Secondary Temperature Anisotropies

On small scales, inhomogeneities in the density, ionized fraction, and velocity field combine to produce temperature anisotropies during reionization. These anisotropies are referred to as the *kinetic Sunyaev-Zel'dovich (kSZ) effect*, which broadly encompasses two distinct physical components, the *Ostriker-Vishniac effect* and the *patchy reionization signal*.

Let us first consider the general expression for the CMB temperature along a line of sight $\hat{\mathbf{n}}$,[8]

$$\frac{\Delta T_\gamma(\hat{\mathbf{n}})}{T_\gamma} = \int d\eta\, e^{-\tau_{\rm es}(\eta)} a n_e(\eta) \sigma_T \frac{\hat{\mathbf{n}} \cdot \mathbf{u}(\eta)}{c}, \tag{13.3}$$

where $\eta = \int_0^t dt'/a(t')$ is the conformal time, $\mathbf{u}$ is the peculiar velocity of the ionized gas, and $\tau_{\rm es}(\eta)$ is the optical depth between the observer and a conformal time η. (The extra factor a occurs because $c\,dt = a\,d\eta$.) Perturbations in the temperature are then sourced by the product of the peculiar velocity along the line of sight and the ionized gas density; we define $\mathbf{q} = \mathbf{u}(1 + \delta_b + \delta_x)$ for convenience. Here, we must be careful to include the baryonic density fluctuation (rather than cold dark matter) because it is this material that scatters the photons.

Suppose then that we work to linear order, so that $\mathbf{q} \approx \mathbf{u}$ (since the peculiar velocity is itself a first-order quantity). Because the Fourier transform of this velocity is parallel to the wave vector $\mathbf{k}$ (see equation ~~2.15~~), this implies that only

2.14 / from equation renumbering

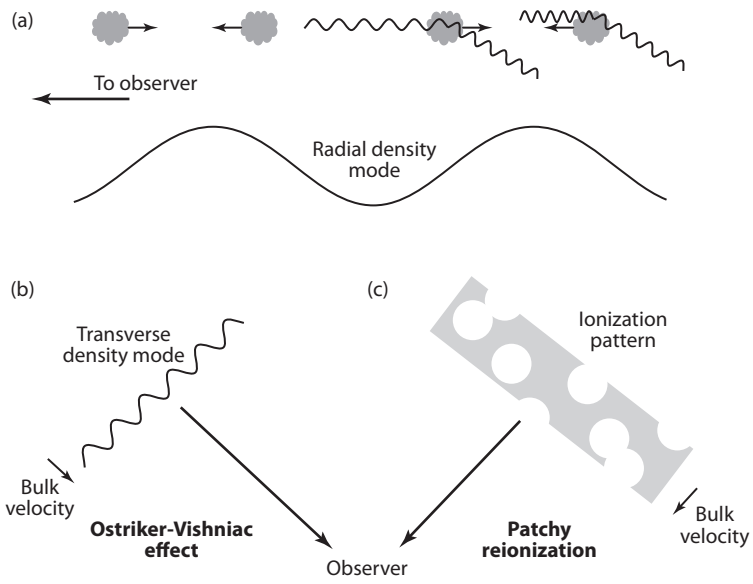


Figure 13.4 Diagram of the kinetic Sunyaev-Zel'dovich effect. *(a)* We imagine a plane wave perturbation oriented along the line of sight. The density perturbations generate peculiar velocities as matter falls into the overdense regions. CMB photons that scatter on the far side of these perturbations get blueshifted (on average), but photons that scatter on the near side get redshifted. To linear order, there is a net cancellation of the two frequency shifts. *(b)* However, at higher order the Doppler shifts do not cancel out, so long as another density mode affects the scattering. Here, we imagine a region with a net velocity toward the observer and a density mode *transverse* to the line of sight. Overdensities across the sky generate a stronger blueshift than the underdensities owing to their larger optical depth. On small scales, these do not cancel with other structures along the line of sight. This density modulation is called the *Ostriker-Vishniac effect*. *(c)* Alternatively, we imagine another region with a bulk velocity directed toward the observer, in which the ionized fraction varies spatially owing to discrete H II regions during reionization. This modulation is associated with *patchy reionization*. Both nonlinear effects *(b)* and *(c)* source CMB temperature fluctuations on small angular scales.

those modes along the line of sight contribute. But for such waves oriented along the line of sight, their troughs and crests will (nearly) cancel once we perform the integration, especially on small angular scales, where there are many such troughs and crests. This is illustrated in Figure 13.4(a). (Modes transverse to the line of sight do not suffer such a cancellation—but, of course, they do not cause any Doppler shift either.)

Thus, angular correlations can be generated only by modes perpendicular to the line of sight (which themselves would not contribute if only $\mathbf{u}$ appeared—it is the *nonlinear* terms $\delta_b\mathbf{u}$ and $\delta_x\mathbf{u}$ that generate anisotropies). In that case, the

components of $\mathbf{q}$ parallel to $\mathbf{k}$ also do not contribute to the anisotropy because they are small after projection. The anisotropy is therefore generated only by the component of $\mathbf{q}$ perpendicular to the wave vector, which we call $\mathbf{q}_\perp$. After projection on the sky, the resulting angular fluctuation power spectrum is[9]

$$C_\ell = (\sigma_T \bar{n}_e^c)^2 \int \frac{d\eta}{r^2} W(\eta)^2 P_{q\perp}(\ell/r, \eta), \tag{13.4}$$

where $\bar{n}_e^c$ is the total comoving electron density, r is the comoving distance from the observer to a conformal time η, $W = \bar{x}_i e^{-\tau_{es}}/a^2$, and $P_{q\perp}$ is the three-dimensional power spectrum of the projection $\mathbf{q}_\perp$. This integral picks out the physical scales corresponding to a given observed multipole moment, weighting the contribution from each redshift by the factor W^2/r^2, which includes both the IGM ionized fraction (or the fraction of matter that can actually scatter CMB photons at the relevant redshift) and the effect of subsequent scattering that washes out the secondary anisotropies (via the exponential). We illustrate the physics of this process in Figure 13.4.

The power spectrum $P_{q\perp}$ involves four-point functions (or correlations between four quantities) in many possible combinations, some of which are negligible. For example, in practice these four-point functions factor into pairs of normal power spectra, because the "connected" higher-order correlations vanish for Gaussian random fields. Moreover, terms like $P_{\delta_b v}$ (i.e., the cross–power spectrum between baryon density and velocity) can be ignored because of the scale mismatch between these two quantities (recall that in linear theory, $v \propto \delta/k$, so it is driven by large-scale modes—while the δ fluctuations of interest occur only on small scales).

In most models, the dominant contribution to $P_{q\perp}$ is[10]

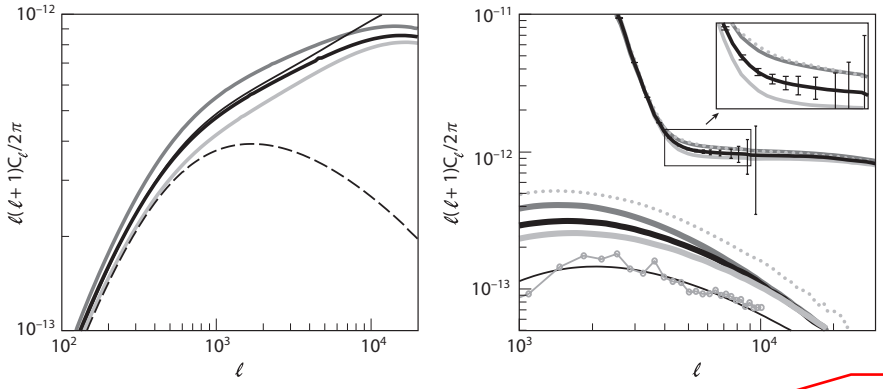
$$P_{\rm OV} = \frac{1}{3} v_{\rm rms}^2 P_{\delta_b \delta_b}, \tag{13.5}$$

or the Ostriker-Vishniac effect, which arises from scattering off ionized clouds with bulk motions. The rms velocity is given by

$$v_{\rm rms}^2 = \int dk \frac{k^2}{2\pi^2} P_{vv}(k), \tag{13.6}$$

where P_{vv} is the power spectrum of the velocity (cf. equation 2.24). Because this effect is most important on small scales, one must usually use the nonlinear density power spectrum (filtered appropriately for small-scale baryonic structure) to evaluate the $P_{\delta_b \delta_b}$ contribution—either through numerical simulations or an approximation like the halo model.

The Ostriker-Vishniac effect has contributions from all redshifts at which the IGM (or even gas near or within galaxies) is ionized, and it is strongest at lower redshifts, where both the rms velocity field and the density fluctuations are most significant. Its total amplitude does, however, depend on when reionization began, and its shape depends very slightly on this as well (both because of the changing angular diameter distance and the evolving characteristic scale of structure formation).



from bottom to top / AA

Figure 13.5 *Left:* Angular power spectrum of the Ostriker-Vishniac effect on the CMB. The thick ~~dashed,~~ solid ~~and dot-dashed~~ curves take models in which reionization ends at $z_{\rm reion} = 8$, 12, and 18, ~~respectively~~. The thin solid curve shows the signal without any baryonic filtering, while the thin dashed curve shows the signal assuming only a linear theory density field. *Right:* The patchy reionization signal from the same models (lower left, thick curves) and the total CMB anisotropy power spectrum (at upper right, including the primordial, lensing, and Ostriker-Vishniac components as well); note the change in the scale of the vertical axis on this panel. In addition, the dotted curve shows the patchy signal from a model with extended reionization. The open circles show the estimate from a seminumerical simulation of reionization, with the thin line a corresponding analytic estimate with the same reionization history; the two match very well. The error bars at upper right are representative of existing instruments, assuming perfect foreground removal. McQuinn, M., et al., *Astrophys. J.* **630**, 643 (2005). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

The left panel of Figure 13.5 shows some example angular power spectra of the Ostriker-Vishniac effect in three different models of reionization; in each case $\bar{x}_i \propto f_{\rm coll}$ and the models are calibrated so that reionization ends at $z_{\rm reion} = 8$, 12, and 18 (thick dashed, solid, and dot-dashed curves, respectively). Note that even with these rather different histories, the Ostriker-Vishniac signal changes by only ∼10%. Thus, deducing information about reionization from this component of the CMB anisotropies will be challenging, requiring very careful modeling of the dominant lower-redshift contributions.

The other two terms that contribute to $P_{q\perp}$ involve integrals over $P_{vv}P_{\delta_x\delta_x}$ and $P_{vv}P_{\delta_b\delta_x}$, which are known as the "patchy reionization" contributions.[11] Because these involve fluctuations in the ionized fraction, they are relevant only during the reionization era itself and so better isolate that period's properties. Physically, they originate from the peculiar velocities of the ionized bubbles that appeared during reionization, which have biased velocities relative to the background. They can actually be estimated analytically using the simple tools describing the statistical properties of the ionization field developed from the excursion set model of reionization in §9.7.

The patchy reionization contribution therefore depends both on the structure of the ionized bubbles and their relation to the density field; the right panel of Figure 13.5 shows some example models. The thick ~~dashed~~, solid, and dot-~~dashed~~ curves at bottom left show this component of the signal for the same models as in the left panel. In all cases, the signal peaks at $\ell \sim 2000$, which is simply the angular scale corresponding to the projected physical sizes of the bubbles midway through reionization (where the patchiness peaks). The different amplitudes of the contribution come not from differences in the reionization models—$n_b(m)$ is very similar in all three of these—but because of the different ionized gas densities during reionization.

Actually, the *duration* of reionization is the most important factor determining the amplitude of the patchy signal: the longer the contrast between ionized and neutral gas persists, the longer the patchiness continues to generate CMB fluctuations. The thick dotted curve illustrates this condition: it takes a model in which reionization lasts roughly twice as long as in the solid curve (but ends at the same time). Note that the patchy signal increases by nearly a factor of 2. Thus, CMB secondary anisotropies provide a tool for studying not only the timing of reionization but also its duration.

The upper right set of curves in the right panel of Figure 13.5 show the sum of the Ostriker-Vishniac, patchy reionization, and primordial anisotropies (including lensing); the inset zooms in on the most useful part of the spectrum for studying the cosmic dawn. In this regime, the primordial anisotropies die off very quickly owing to diffusion (Silk damping) inside the recombination surface; nevertheless, they are still very large near the peak of the patchy reionization contribution. Instead, it is the tail of this distribution, together with the Ostriker-Vishniac effect generated by all structure past reionization, that can be separated. The error bars here show estimates for ongoing ground-based CMB surveys, whose early results are encouraging.[12] Despite the apparently modest differences among the models, such an experiment can still easily distinguish them.

However, the primary challenge to making these measurements are contaminants: both lower-redshift point sources and other CMB secondary anisotropies pose substantial problems. By far the strongest secondary at these angular scales is the *thermal Sunyaev-Zel'dovich effect*, which describes the frequency shift undergone by CMB photons scattering inside hot gas (in particular nearby galaxy clusters). In principle, because this scattering process is frequency dependent (and disappears entirely at 218 GHz), it can be separated from other anisotropies (including the kinetic Sunyaev-Zel'dovich effect). Constraints on the cosmic dawn from these secondaries will rely on accurate removal of these "foregrounds" as well as accurate modeling of the Ostriker-Vishniac effect at lower redshifts. A particularly insidious problem is the correlation between the point sources and the thermal Sunyaev-Zel'dovich effect.

In principle, sources from the cosmic dawn may also contribute to this thermal Sunyaev-Zel'dovich signal, which arises when hot electrons undergo inverse-Compton scattering off the CMB photons, and transfer energy to the photon field. The magnitude of this effect is parameterized by the Compton-y

parameter, which is the typical energy transfer per scattering times the number of scatterings. It also measures the resulting spectral distortion in a blackbody spectrum: in the Rayleigh-Jeans limit, the temperature distortion is $\Delta T_\gamma / T_\gamma = -2y$. The contribution from a volume element with electron scattering optical depth $d\tau_{\rm es}$ is

$$dy = \frac{k_B(T_e - T_\gamma)}{m_e c^2} d\tau_{\rm es}. \tag{13.7}$$

Thus, the total y-distortion depends on both the ionization and thermal histories of the gas; importantly, however, it includes both gas *inside* and *outside* galaxies.

The sources that reionized the Universe may very well make a nonnegligible contribution to the overall y-distortion, owing to the supernovae that inevitably accompany the massive stars able to ionize the IGM. Even though these supernovae may remain confined to the ISM of their host galaxies, they still likely lose a substantial fraction of their energy to the CMB, inducing a y-distortion. Let us begin by computing the energy injection per baryon from such explosions. We write $\omega_{\rm SN}$ for the supernova energy produced per solar mass in stars formed; this is $\sim 10^{49}$ erg $M_\odot^{-1}$ for typical stellar IMFs (see §6.4.1). We then write the fraction of baryonic mass in stars, $f_\star f_{\rm coll}$ according to our usual definitions, in terms of the number of ionizing photons produced per baryon in the Universe, $Q_{\rm H\ II}$, and the number produced per baryon inside stars, $N_{\rm ion}$. Then, the available thermal energy per baryon is

$$\frac{\epsilon_{\rm SN}}{\bar{n}_b} \sim 10 Q_{\rm H\ II} \left(\frac{400}{N_{\rm ion} f_{\rm esc}} \frac{\omega_{\rm SN}}{10^{49}\,{\rm erg}/M_\odot} \right) {\rm eV}. \tag{13.8}$$

Interestingly, this energy produced in supernovae is quite close to the amount actually needed to ionize the IGM ($\sim$13.6 eV per baryon), though the amount injected into the CMB may be much smaller.

Now, let us assume that this energy injection occurs at some redshift $z_{\rm SN}$. The CMB spectral distortion satisfies

$$y = -\frac{1}{2}\frac{\Delta T_\gamma}{T_\gamma} \sim -\frac{1}{8}\frac{\Delta U_\gamma}{U_\gamma} \sim 10^{-6} f_{\rm comp} \left(\frac{10}{1+z_{\rm SN}} \right) \left(\frac{\epsilon_{\rm SN}/\bar{n}_b}{20\ {\rm eV}} \right) \tag{13.9}$$

where $f_{\rm comp}$ is the fraction of the supernova energy that is actually injected into the CMB. This is a reasonably large number: the current observational limit, from the FIRAS instrument on the Cosmic Background Explorer (COBE) is $y \leq 1.5 \times 10^{-5}$,[13] and models that predict signals from lower redshifts are just a factor of about three times larger than equation (13.9). (However, the photo-heating accompanying reionization does *not* make a substantial contribution: typically just a fraction of an electron-volt is injected to the CMB per baryon, which produces a very small signal compared with the much hotter gas at lower redshifts.)

Such an observable signal hinges on the injection of the supernova energy into the CMB, so that $f_{\rm comp} \sim 1$. As we saw in §6.4.2, many other processes

move 13 as superscript of `is'
to avoid confusion with power-law 5
/AA

help cool the remnants, especially if they remain confined to their galaxies. The most important is radiative cooling in the dense shell plowing through the IGM (or ISM). Simple estimates suggest that the energy contained in the explosion blast wave must be very large and that the remnant must spend the bulk of its time plowing through gas near the IGM density for more than just a few percent of the energy to be lost to the CMB.[14] This would require a strong superwind to escape into the IGM and would have many other important consequences, such as metal enrichment and mixing of the IGM gas (see §6.4 and §6.5.2).

If any of these mechanisms do cause a substantial y-distortion, the strong clustering of early stellar sources would also induce substantial angular fluctuations in the CMB temperature field through the thermal Sunyaev-Zel'dovich effect, which would then also be observable once the contribution from nearby hot galaxy clusters was subtracted. Again, the modeling of lower-redshift contaminants, which depends on uncertain factors like cluster cooling and turbulence, AGN feedback, and the poorly understood properties of gas inside small galaxy groups, will be crucial to disentangling any possible high-redshift contribution.

13.2 Diffuse Backgrounds from the Cosmic Dawn

Although many of the large telescopes planned for the next decades can study individual high-redshift galaxies and quasars in exquisite detail, one can also learn a great deal about these objects—even without resolving them—by studying the *integrated* radiation backgrounds generated by such sources. If one does not attempt to identify individual galaxies, the telescope requirements are more modest, and the characteristics of the *entire* galaxy population can be measured. Crucially, unlike in a traditional galaxy survey, measures of the the integrated emission are sensitive even to extremely faint galaxies—which may dominate the *cumulative* luminosity density of the Universe if the luminosity function is steep.

Of course, these benefits come with a price—diffuse backgrounds are much more difficult to interpret in the presence of other astronomical (or terrestrial) backgrounds that they overlap. Much as with the spin-flip background, the observational challenge is typically to extract subtle cosmological information from a much larger net signal. This task is easiest for a background generated by an emission line, because it then has both angular and frequency structure that clearly reflects the source population. Broadband backgrounds can be distinguished only by resolving other possible contaminants (as in the X-ray background discussed in §9.8.1); we therefore focus primarily on line backgrounds here.

The first interesting aspect of such backgrounds is their amplitude as a function of redshift, which provides a measure of the total emissivity in this radiation mechanism through the cosmic dawn. For an emission line, where each observed frequency corresponds to a different distance, such a background is even more useful because it allows us to measure redshift evolution.

Let us assume that a source population has a specific emissivity $\epsilon(\nu, \mathbf{r}, z)$ (with units of energy per time per frequency per comoving volume). Neglecting intervening absorption, the observed specific intensity I_ν at a frequency $\nu_{\rm obs}$ along a line of sight $\hat{\mathbf{n}}$ is the integral of the emissivity (cf. equation 4.43),

$$\nu_{\rm obs} I_{\nu_{\rm obs}} = \frac{c}{4\pi} \int dz\, \nu(z) \frac{\epsilon[\nu(z), \hat{\mathbf{n}} r(z), z]}{H(z)(1+z)^2}, \tag{13.10}$$

where $\nu(z) = \nu_{\rm obs}(1+z)$ is the emission frequency at a redshift z, and $r(z)$ is the comoving distance along the line of sight to the source. If ϵ_ν extends over a wide frequency range, $I_{\nu_{\rm obs}}$ will therefore sample a wide redshift range. But if ϵ_i describes a line with rest wavelength ν_i, then the observed intensity will sample only a specific redshift $z_{\rm obs} = \nu_i/\nu_{\rm obs} - 1$. As with the spin-flip background, we can use this relation between observed frequency and distance to map the structure of the emission in three dimensions.

Without resolving individual sources, background measurements will be sensitive to large-scale fluctuations in the emissivity, which we can easily parameterize with the galaxy power spectrum, either taken from simulations or computed with the halo model (see §3.6.1). For example, let us assume that the luminosity in line i of a halo with mass m is $L_i(m)$ and that a duty-cycle fraction $f_{\rm duty}$ (which may also be a function of mass) of dark matter halos with $m > m_{\rm min}$ emit in this line. Then, the mean emissivity in the line is

$$\langle \epsilon_i \rangle(z) = \int_{m_{\rm min}}^{\infty} dm\, L_i(m)\, f_{\rm duty}(m) n(m). \tag{13.11}$$

Note that this integral has units of energy per (comoving) volume per second and is not a specific emissivity; rather, it includes all the emissions in the line.

The fluctuations also trace the population of massive halos, so it is natural to use the halo model to estimate them. However, for a diffuse background originating from unresolved sources, we are typically not concerned with the signal structure on scales below the typical halo's virial radius, which is in any case likely to be simply a set of point sources generated by each galaxy residing in the halo. We can therefore treat each halo as a single point source and ignore the one-halo term.

In this case, the power spectrum of fractional emissivity fluctuations is $P_i(k) \approx P_i^{2h}(k)$, with

$$P_i^{2h}(k) = P_{\rm lin}(k) \left[\int dm \frac{L_i(m)\, f_{\rm duty}(m) n(m)}{\langle \epsilon_i \rangle} b_{\rm eff}(k|m) \right]^2, \tag{13.12}$$

where we have used the effective scale-dependent bias that incorporates nonlinear effects in the two-halo term (see §3.6.3) and assumed that we are on sufficiently large scales that $u_{\rm gal}(k|m) \approx 1$ for all the halos of interest. Mapping the spatial fluctuations in this background will therefore inform us about the number densities of these sources together with their scale-dependent bias; in principle, the latter is separable because of its special shape, so these different physical effects can be measured. (If the observations extend to scales fine

enough to allow the one-halo term to be measured as well, they will allow a similar separation.)

Even though the one-halo term itself is significant only on small scales, the shot noise arising from the finite number of sources (see §3.6.4 and §10.4.1) can be significant, because halos massive enough to host galaxies were still relatively rare at these early times. Shot noise can be especially important if a small duty cycle limits the fraction of time for which any individual halo is luminous or if only very massive halos emit strongly in the relevant line. This shot-noise term produces a white-noise spectrum (i.e., independent of k) with the amplitude of the emissivity power spectrum equal to

$$P_{\mathrm{shot},i}(k) = \int_{m_{\mathrm{min}}}^{\infty} dm \frac{L_i^2(m)}{\langle \epsilon_i \rangle^2} f_{\mathrm{duty}}(m) n(m). \tag{13.13}$$

With equation (13.11) for the mean emissivity, the shot-noise term decreases with increasing source density, just as in the simple counting case (see §10.4.1). However, in contrast with our earlier discussion of shot noise (see equation 10.22), in which we were concerned only with counting galaxies, for a diffuse background the contribution from different sources to the observed signal is weighted by their luminosity—thus, this expression has a factor of L_i^2 inside the integral.

The shot noise is therefore dominated by the most luminous sources, and one can often substantially decrease its amplitude by removing those bright sources that can be individually detected. In that case the remaining diffuse background is the sum of the emissions from galaxies below the detection limit of the survey—which extends the "dynamic range" of measurements. Alternatively, the shot-noise term can often be removed statistically, because its shape (white noise) is precisely defined.

Note that we have expressed these fluctuation spectra in their *fractional* forms —that is, P_i has units of volume, independent of the emissivity. This means that they apply equally well to the emissivity or to the observed intensity—for example, one simply multiplies by the mean observed intensity to recover the latter in dimensional form.

13.2.1 The Near-Infrared Background

As an important example, we consider the integrated background from UV photons emitted during the cosmic dawn. This background has two principal components: first, a broadband background from stellar continua, and second, line and continuum components from the reprocessing of ionizing photons—which principally result in Lyman-α photons (see chapter 11). The latter is the most useful, because the broadband component involves a mix of many emission redshifts for any given observed frequency, making tomography impossible.

To estimate the monopole amplitude of this background and its spectrum, we must estimate the contribution of all these different processes. Assuming that the background is generated entirely by star formation, a convenient

parameterization is

$$\epsilon(\nu, z) = \dot{\rho}_\star(z) c^2 \sum_i \langle f_\nu^i \rangle, \tag{13.14}$$

where $\dot{\rho}_\star$ is the formation rate of stellar mass per comoving volume, the sum extends over all radiation processes (labeled by i), and $\langle f_\nu^i \rangle$ is the fraction of the stellar rest mass energy that is released in process i in the frequency interval $(\nu, \nu + d\nu)$ (for stars, it must therefore be averaged over both their IMF and their lifetime). The nuclear burning efficiency for stars implies that $\nu \langle f_\nu^i \rangle \sim 10^{-3}$. With this estimate, and assuming high redshifts for the integrand, equation (13.10) becomes[15]

$$\nu_{\rm obs} I_{\nu_{\rm obs}} \approx 11 \text{ nW m}^{-2} \text{ sr}^{-1} \int \frac{dz}{1+z} \dot{\rho}_\star(z) \left(\frac{10}{1+z}\right)^{5/2} \sum_i \left(\frac{\nu(z) \langle f_{\nu(z)}^i \rangle}{10^{-3}}\right), \tag{13.15}$$

where $\dot{\rho}_\star$ is measured in units of $\text{M}_\odot \text{ yr}^{-1} \text{ Mpc}^{-3}$.

The following processes contribute to this background:

- The stellar continuum, which (below the Lyman edge) can be roughly approximated as a blackbody with some effective temperature $T_{\rm eff}$ that is a function of stellar mass.
- Free-free and free-bound emission from H II regions. The total luminosity of this component may be written as

eliminate overdot / AA

$$L_\nu^{\rm ff,fb} = \frac{\epsilon_\nu^{\rm ff,fb} \dot{Q}_i \dot{M}_\star}{n_e n_p \alpha_B}, \tag{13.16}$$

  where $\epsilon_\nu^{\rm ff,fb}$ is the volume emissivities of these processes, $\dot{Q}_i$ is the production rate of hydrogen-ionizing photons per unit star formation rate, and $\dot{Q}_i \dot{M}_\star / n_e n_p \alpha_B$ is simply the volume of the ionized regions, assuming the Strömgren sphere limit. Because these processes are generated by collisions of electrons and protons, the emissivity is

$$\epsilon_\nu^{\rm ff,fb} = 4\pi n_e n_p g_{\rm eff} \frac{e^{-h\nu/k_B T}}{T}, \tag{13.17}$$

  where the "Gaunt factor" $g_{\rm eff} \sim 1$ depends weakly on temperature and density.[16] Thus, the *total* luminosity is independent of the local density. This means that for the purposes of estimating the diffuse background, we do not need to worry about the internal structures of the galaxies, or even whether the ionizing photons escape into the IGM: that will clearly affect the small-scale spatial distribution of the photons, but it has no effect on the overall luminosity, as long as the H II regions are unresolved and in ionization equilibrium.
- Recombination line emission—and in particular the Lyman-α line (higher Lyman-series photons are absorbed and cascade to either Lyman-α or the two-photon continuum discussed next). In §11.1, we saw that the Lyman-α profiles (both spatial and spectral) of individual galaxies depend strongly

on the structure and physics of the galaxies and their local environments. However, the integrated emission is much simpler to estimate; assuming only that the mean free path of an ionizing photon is much shorter than the Hubble length, and that ionization equilibrium applies, the net luminosity of Lyman-α photons is simply (cf. equation 11.2)

$$L_{\mathrm{Ly}\alpha}^{\mathrm{db}} = \frac{2}{3} T_{\mathrm{Ly}\alpha}^{\mathrm{ISM}} \dot{\overline{Q}}_i h\nu_\alpha \dot{M}_\star. \tag{13.18}$$

eliminate overdot / AA

Here, $T_{\mathrm{Ly}\alpha}^{\mathrm{ISM}}$ accounts for absorption—and subsequent destruction—of Lyman-α photons by dust; note that there is no corresponding factor for absorption of the Lyman-α photons inside the IGM, because those photons are scattered but not destroyed. As with the free-free and free-bound processes, there is no distinction between ionizations that occur inside a galaxy and in the IGM, for both can produce Lyman-α photons through recombinations (though we do assume here that recombinations instantaneously follow ionizations, which is not accurate at low IGM densities). We ignore the Balmer series or longer-wavelength transitions because they carry much less energy than the Lyman-α line.

- Two-photon emission from (forbidden) decays between the $2S$ and $1S$ levels of hydrogen. The former level can be populated by radiative cascades following recombinations. The luminosity of this process is again proportional to the rate of ionizing photon production,

$$L_\nu^{2\gamma} = \frac{2h\nu}{\nu_\alpha}(1 - f_{\mathrm{Ly}\alpha}) P(\nu/\nu_\alpha) \dot{\overline{Q}}_i \dot{M}_\star, \tag{13.19}$$

where $f_{\mathrm{Ly}\alpha} \approx 0.64$ is the fraction of cascades that result in Lyman-α photons (here we have assumed efficient mixing of the angular momentum states), the factor 2 appears to account for the two photons produced in each decay, and $P(x)dx$ is the normalized probability per two-photon decay of obtaining one photon in the range $dx = d\nu/\nu_\alpha$.[17]

Figure 13.6 shows some example (monopole) spectra containing all these processes; they are normalized to the overall star formation rate density $\dot{\rho}_\star$. These examples include only stars from $z = 7$–15. In the left and right panels ~~varies~~ the metallicity varies (which has a significant effect on the rate of ionizing photon production); within each panel, the different line styles take different prescriptions for the IMF. The solid line corresponds to a standard Salpeter IMF, with the number of stars forming per unit mass $\propto m^{-2.35}$. The dashed line refers to a Larson IMF, with the mass spectrum $\propto m_\star^{-1}(1 + m_\star/m_c)^{-1.35}$, where the characteristic mass is $m_c = 50\,M_\odot$. Finally, in the left-hand panel the ~~double~~-dot-dashed curve takes a flat distribution by mass over the range 100–500 $M_\odot$.

PE

AA

The curves in each panel show the contributions of different processes to the overall spectrum. The uppermost is of course the total background per unit star formation rate. The straight line peaking at $\sim 1\,\mu$m shows the contribution from Lyman-α photons; note that the shape is simply a consequence of the assumption of a constant star formation rate and is not a robust prediction.

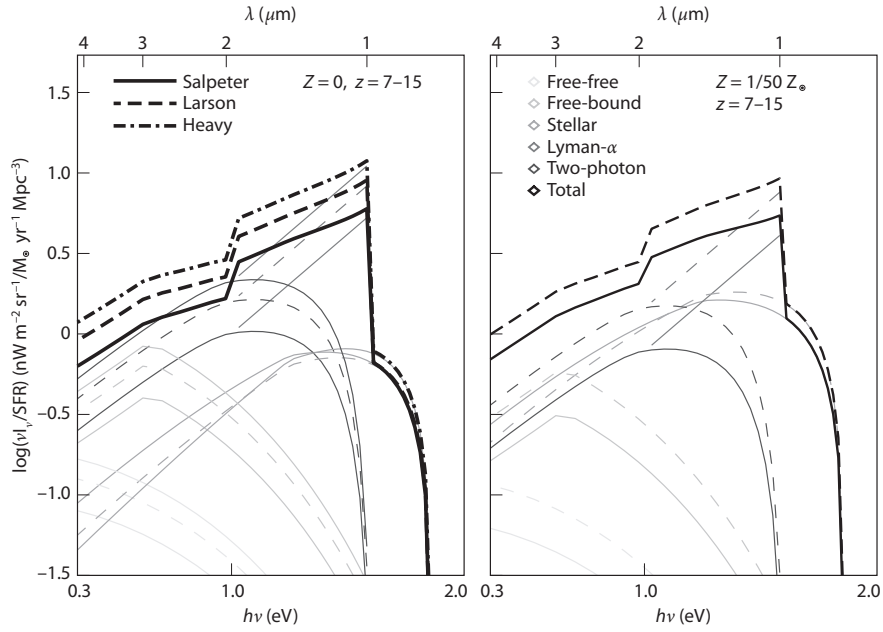


Figure 13.6 Near-IR background spectrum (see *Color Plate 31* for a color version of this figure). This example includes star formation in the range $z = 7$–15, which dominates the background between 1 and 2 μm. The left panel assumes metal-free stars, while the right panel takes a metallicity $Z = 0.02\ Z_\odot$. The different line styles show different IMFs (see text). The different curves show the total emission, the Lyman-α contribution, the stellar continua, the two-photon continua, the free-bound emission, and finally the free-free emission (from top to bottom at 1 μm). Fernandez, E. R., & Komatsu, E., *Astrophys. J.* **646**, 703 (2006). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

The next strongest process, which peaks near 1 μm as well, is the stellar continuum; the curves here continue through the Lyman series without taking into account the sawtooth IGM absorption of these photons (see Figure 6.2). The curve peaking at somewhat lower energy, but with a comparable amplitude, comes from the two-photon decays: note that this is well below the Lyman-α peak not because significantly less energy goes into this process but because it is distributed over a wide frequency interval. Finally, the lowest amplitude curves (visible at lower left) show the contribution from free-bound and free-free emission, neither of which is significant.

Figure 13.6 shows several interesting points. First, in the most interesting wavelength range ($\sim$1–2 μm here) the background is usually dominated by the Lyman-α photons—especially for the (hotter) metal-free stars—because this line contains a significant fraction of the entire ionizing luminosity of the starlight. Nevertheless, other processes provide nontrivial corrections at higher

wavelengths. Note that stars at higher redshifts do not significantly affect the background in this band either, because their Lyman-α emission lies at longer wavelengths and (unless the comoving star formation rate becomes much higher at high redshifts) the stellar continua are quite weak. The near-IR background therefore offers a relatively clean probe of the ionizing photon budget during the bulk of the cosmic dawn era.

Unfortunately, measuring this signal is extremely difficult: like the spin-flip background, it suffers severe contamination from "local" sources, IR emission from lower-redshift galaxies and zodiacal light from our own solar system, which is roughly three times the expected signal. This background arises from dust particles inhabiting the ecliptic plane that scatter sunlight and has proved very difficult to model with sufficient precision to extract the high-redshift signal reliably.

Fortunately, as with the spin-flip background, *fluctuations* in the near-IR background light may be easier to detect than this monopole spectrum, because variations in the foreground contaminants have different spatial and spectral structure than the high-redshift light. In fact, the current best estimates of the monopole background come from measurements of the fluctuations themselves; they indicate that the excess over the known backgrounds cannot be much larger than $\sim 1\,\mathrm{nW\,m^{-2}\,sr^{-1}}$. This value provides an interesting limit on the cosmic star formation rate at $z \sim 10$.[18]

However, for higher-frequency diffuse backgrounds like this one, it is very difficult to recover fluctuations along the line of sight, as that requires an integral field spectrograph with sufficiently high spectral resolution to separate the features. To date, no such instruments are available over the wide fields of view necessary to measure this background in the near-IR. Instead, these backgrounds are integrated over a finite frequency interval, and, typically, the angular fluctuations are measured (just as in the CMB) rather than the three-dimensional structure. This makes the foreground subtraction somewhat more difficult (unless several contiguous filters are used), because the spectral information is lost. Most often, the subtraction relies on modeling of the foreground structures.

In this case, we measure the band-averaged intensity $I(\hat{\mathbf{n}})$,

$$I(\hat{\mathbf{n}}) = \frac{c}{4\pi}\int dz\,\frac{\tilde{\epsilon}[\hat{\mathbf{n}}r(z), z]}{H(z)(1+z)^2}, \tag{13.20}$$

where $\tilde{\epsilon}$ is the integral of the comoving specific volume emissivity over the emission frequency range corresponding to the observed band.

To construct the angular fluctuation spectrum, we take the spherical harmonic transform of equation (13.20). We begin with Rayleigh's formula for the spherical decomposition of a plane wave,

$$e^{-i\mathbf{k}\cdot\mathbf{x}} = 4\pi\sum_{\ell m}(-i)^\ell j_\ell(kx)Y^*_{\ell m}(\hat{\mathbf{k}})Y_{\ell m}(\hat{\mathbf{n}}), \tag{13.21}$$

where j_ℓ is the spherical Bessel function of order ℓ, and $Y_{\ell m}$ are the spherical harmonics. As a reminder, ℓ fixes the angular scale of variations in the function,

and m chooses among the possible configurations on that scale. This decomposition determines the expansion coefficients of

should be script ell (as in rest of equation) / AA

$$I(\hat{\mathbf{n}}) = \sum_{lm} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}) \tag{13.22}$$

in terms of the three-dimensional Fourier transform of the emissivity, $\tilde{\epsilon}(\mathbf{k}, z)$,

$$a_{\ell m} = c(-i)^{\ell} \int \frac{dz}{H(z)(1+z)^2} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \tilde{\epsilon}(\mathbf{k}, z)\, j_{\ell}[kr(z)] Y^*_{\ell m}(\hat{\mathbf{k}}). \tag{13.23}$$

The angular power spectrum is usually expressed as the ensemble average of the spherical harmonic coefficients, $C_{\ell} = \langle |a_{\ell m}|^2 \rangle$. The ensemble average acts on the two factors of the (band-averaged) emissivity to give the three-dimensional power spectrum of the emissivity, $P_{\tilde{\epsilon}}$ (which we define relative to the mean emissivity, $\langle \tilde{\epsilon}(z) \rangle$, so that $P_{\tilde{\epsilon}}$ has units of volume, just like the matter power spectrum). Thus, the angular power spectrum is

$$C_{\ell} = \frac{c^2}{8\pi^3} \int \frac{dz}{H(z)(1+z)^2} \int \frac{dz'}{H(z')(1+z')^2} \times \int k^2 dk \langle \tilde{\epsilon}(z) \rangle^2 P_{\tilde{\epsilon}}(k, z)\, j_{\ell}[kr(z)]\, j_{\ell}[kr(z')]. \tag{13.24}$$

This form has units of (flux per solid angle) squared, because we have integrated the flux density over a narrow band.

Fortunately, the inner integral can be simplified in the small-angle limit ($\ell \gg 1$). The integral of a product of two spherical Bessel functions is

$$\int k^2 dk\, j_{\ell}(kx)\, j_{\ell}(kx') = \frac{\pi}{2} \frac{\delta(x - x')}{x^2}. \tag{13.25}$$

In the large ℓ limit, j_{ℓ} oscillates very rapidly, and the integral is dominated by $k \approx \ell/r$. Thus, provided that $P_{\tilde{\epsilon}}$ is a slowly varying function, we can extract it from the integral and evaluate its argument at $k = \ell/r$. This *Limber's approximation* allows us to write

$$C_{\ell} \approx \frac{c}{(4\pi)^2} \int \frac{dz}{H(z) r^2(z) (1+z)^4} \langle \tilde{\epsilon}(z) \rangle^2 P_{\tilde{\epsilon}} \left(k = \frac{\ell}{r(z)}, z \right). \tag{13.26}$$

In other words, if we use the simple conversion $\ell \approx kr$, the angular power spectrum is simply the projection of the three-dimensional power spectrum. Note that if the observed band is very thin, so that the line component of the background arises from only a narrow redshift window, we can easily invert the angular power spectrum to obtain the three-dimensional version, at least on scales larger than the width of the redshift window (on smaller scales the angular power is damped by cancellations along the line of sight). However, even with such a narrow window, the broadband component still arises from a wide range of redshifts, so it still requires modeling to invert properly.

The cumulative near-IR background fluctuations are again built from the Lyman-α emission, stellar continua, two-photon continua, and free-free/free-bound emission. However, unlike for the mean signal, the spatial location of

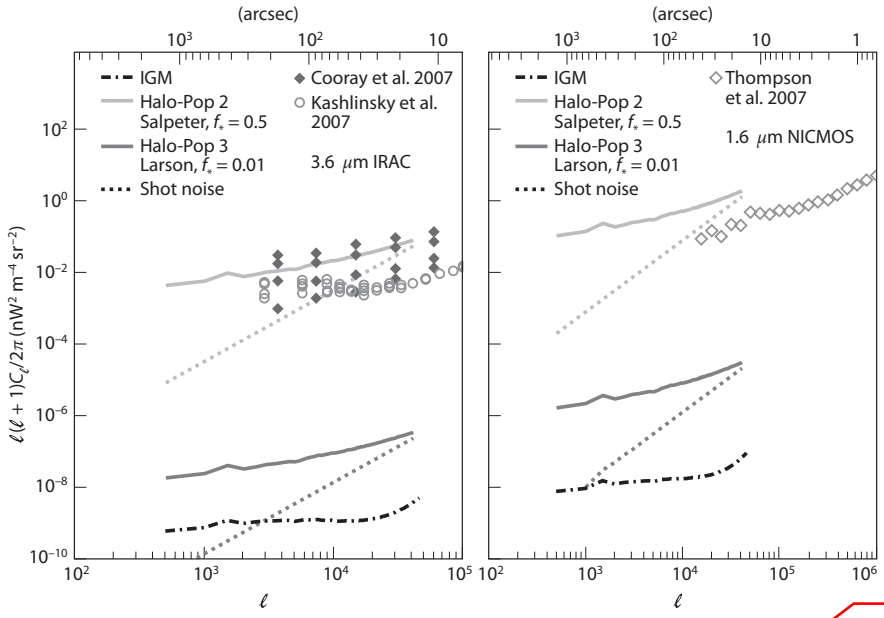


Figure 13.7 Comparison of recent observational data with theoretical predictions (taken from a numerical simulation of reionization) for the angular power spectrum of near-IR background fluctuations. The two panels also show measurements of angular fluctuations in the 3.6-μm IRAC band and the 1.6-μm NICMOS band (see text). In each one, the ~~double-~~dot-dashed curve shows predictions for the IGM emission, while the solid ~~and dashed~~ curves show the range of predictions for the total angular power spectra; these two limiting cases take a light, metal-poor population with high $f_\star$ and small $f_{\rm esc}$ (~~solid~~) and a heavy, metal-free population with low $f_\star$ and high $f_{\rm esc}$ (~~dashed~~ curve), both normalized to have the same overall IGM ionization. The dotted curves show the associated shot-noise terms. Fernandez, E. R., et al., *Astrophys. J.* **710**, 1089 (2010). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

delete x2 / AA

light / AA

dark / AA

by / AAS. /

this emission (within or outside galaxies) is important, because recombinations that occur over the ∼10 Mpc ionized bubbles in the IGM change the spatial scales of the fluctuations (and create such low surface-brightness features that they are all but unobservable in practice). In most circumstances, the fluctuations trace the galactic component and diminish as the escape fraction of UV photons, $f_{\rm esc}$, approaches unity.

Figure 13.7 shows some predictions of this fluctuating background in a range of models of star-forming galaxies during the reionization era. These examples all use a numerical simulation of reionization to predict the signal; the simulation fixes only the *total* ionizing efficiency ζ, so there remains a good deal of freedom in the amplitude of the near-IR background, which has a different dependence on $f_{\rm esc}$ in particular. The two models shown here span the range of possibilities in this simulation (though it is worth noting that there is even

dark solid / AA

light / AA

more variation outside of its particular star formation history). The solid line corresponds to a low-metallicity stellar population with a Salpeter IMF (a mass spectrum $\propto m_\star^{-2.35}$, ranging from 3 to 150 $M_\odot$), a very high $f_\star = 0.5$, and a relatively small $f_{\rm esc} = 0.19$. The dashed line refers to very massive Population III stars (with the Larson IMF described previously and a characteristic mass $m_c = 250\,M_\odot$), a low $f_\star = 0.01$, and an escape fraction $f_{\rm esc} = 1$. Since the latter model has *no* line emission from inside galaxies, the fluctuations seen here are almost entirely due to the stellar continua.

The wide range of amplitudes for the angular power spectrum between these two extreme models illustrates how sensitive the near-IR background is to the parameters of star formation in high-redshift galaxies. In combination with an independent measure of the IGM ionization state, this probe can help break important degeneracies in the ionization efficiency. For a given ζ, the signal is maximized with a large star formation efficiency and small $f_{\rm esc}$ (so that more stars form, increasing both the stellar continuum and the overall production rate of ionizing photons *without* affecting the IGM) and a low-mass stellar population (which also boosts the stellar continuum as compared with the ionization rate).

The shapes of the power spectra are set by a combination of the two-halo term and the shot-noise contribution (which is shown separately by the dotted curves). The latter is characterized by a white-noise power spectrum with $C_\ell \propto \ell^0$. The former mirrors the linear power spectrum multiplied by the bias—in this case, nonlinear corrections are quite significant; otherwise, the power spectrum would have *peaked* at $\ell \sim 10^3$. Instead, in the two-halo term, $\ell(\ell+1)C_\ell \propto \ell^{1/2}$ when the nonlinear bias is included.

Interestingly, current observations are starting to pose interesting limits on these models: the two sets of points in the left-hand panel show independent estimates of the residual fluctuations due to unresolved sources in images from the Infrared Array Camera on the *Spitzer* satellite at 3.6 μm.[19] From our preceding discussion of the mean background, this is most sensitive to Lyman-α at very high redshifts and (most importantly in these models) continuum processes at a range of redshifts. The right panel shows data from the NICMOS camera on the Hubble Space Telescope, which operates at 1.6 μm and so is sensitive to Lyman-α emission at $z \sim 12$ (and continuum processes at all redshifts).[20]

The aim of all these efforts is to find fluctuations near the upper limits of theoretical expectations; however, note that the observed fluctuations may very well include other currently unresolved populations, such as faint low-redshift galaxies. In fact, the solid line in these models corresponds to a *mean* intensity of 15 and 60 nW m^{-2} sr^{-1} in these two bands, well above current limits, which suggests that, in fact, most of the observed fluctuations are due to lower-redshift or local contamination. The dashed line represents a mean background of 0.2–0.8 nW m^{-2} sr^{-1} in these two bands, which can easily be accommodated by estimates of the mean intensity.

The dot-dashed curves in these plots show the prediction for the angular fluctuations generated inside IGM H II regions. Given a particular simulation of

reionization, this component is fixed by the densities and locations of the ionized bubbles. In any case, it is very small—almost always negligible compared with the halo contribution—because of the very low IGM density (and hence recombination and collision rates).

In summary, the near-IR background offers an intriguing view of the evolution of early stellar populations that is highly complementary to other approaches focusing on measuring the ionization state of the IGM or on detecting individual bright objects. The primary challenges to understanding the background are, as in so many other areas, foregrounds: learning how to separate the relatively featureless angular and spectral behavior from other low-redshift contaminants and the zodiacal light. A combination with other, complementary probes may very well prove to be the best way to accomplish this goal.

13.2.2 Diffuse Backgrounds of Radio Lines

Another set of interesting diffuse backgrounds arise from radio and submillimeter lines: the same strong emission lines we discussed in §8.9.4. Two particularly interesting examples are CO, which has a forest of rotational lines with rest frequencies $J\nu_{\rm CO}$ for a transition from excited state J to $J-1$ (here, $\nu_{\rm CO} = 115.3\,{\rm GHz}$), and the fine-structure line of singly ionized carbon C II, which has a rest wavelength of $158\,\mu{\rm m}$ (or frequency 1.9 THz). Unlike Lyman-α, these lines fundamentally trace the *fuel* for star formation rather than the feedback exerted by young stars on their surroundings. [C II] is a common line from neutral gas (because the ionization potential of C I lies below that of H I), and it is an important coolant in the outskirts of gas clouds. CO, however, forms in the molecular complexes out of which stars form, and it is an important coolant in that gas. Table 8.1 lists the most prominent interstellar emission lines in star-forming galaxies, along with their characteristic luminosity per star formation rate [in units of $L_\odot(M_\odot\,{\rm yr}^{-1})^{-1}$].

Some of these lines are particularly strong; for example, the [C II] line may carry as much as a percent of the total luminosity of nearby quiescent galaxies like the Milky Way. Another advantage of these lower-frequency lines, which owing to the cosmological redshift are observed in the centimeter or millimeter range, is the relative ease of building instruments with both large fields of view and good spectral resolution to measure a diffuse background. Such measurements are then ideally suited for cross-correlation with other lines or the spin-flip background (see §13.3).

Moreover, cross-correlation is likely necessary to recover any of these line backgrounds, because a single observed frequency will pick up emission from many different lines at many different redshifts. For example, an observed band around 30 GHz will be sensitive to CO(2–1) at $z = 6.7$ and to CO(1–0) at $z = 2.8$. One can isolate the high-redshift signal by comparing two *different* lines at the proper observed frequencies. For example, CO(1–0) at 10.5 GHz and CO(2–1) at 21 GHz both sample $z = 10$ galaxies. So long as no other emission lines have the same spacing, the cross-correlation between these two measurements will

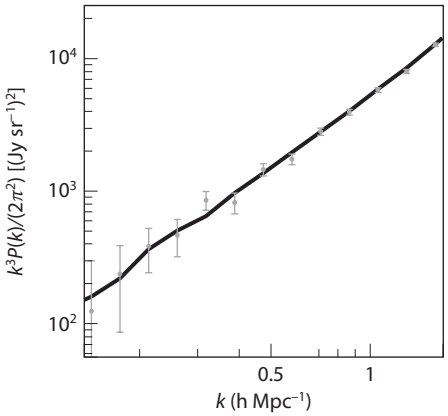


Figure 13.8 The cross–power spectrum of O I(63 μm) and O III(52 μm) at $z = 6$ measured from mock simulation data for a hypothetical IR space telescope (see text). The solid line is the cross–power spectrum measured when only line emission from galaxies in the target lines is included. The points with error bars are the recovered power spectrum when detector noise, contaminating line emission, galaxy continuum emission, dust in our Galaxy, and the CMB are included (see equation 13.29). Visbal, E., Trac, H., & Loeb, A., *J. Cosm. Astropart. Phys.* **11**, 16 (2011). Copyright 2011 by the Institute of Physics.

pick up information from $z = 10$ while eliminating all the contaminants, which simply contribute to the overall noise.

For unresolved point sources emitting in a pair of lines labeled 1 and 2, the cross–power spectrum at a wavenumber k can be approximated as

$$P_{1,2}(\mathbf{k}) = \bar{S}_1 \bar{S}_2 \bar{b}^2 P_{\rm lin}(\mathbf{k}) + P_{\rm shot}, \tag{13.27}$$

where $\bar{S}_1$ and $\bar{S}_2$ are the average fluxes in lines 1 and 2, respectively; $\bar{b}$ is the average bias factor of the sources; $P_{\rm lin}(\mathbf{k})$ is the (linear) matter power spectrum; and $P_{\rm shot}$ is the shot-noise power spectrum due to the discrete nature of galaxies (see equation 13.13). The rms error in the cross–power spectrum at a particular k-mode is given by

$$\delta P_{1,2}^2 = \frac{1}{2}(P_{1,2}^2 + P_{1\rm total} P_{2\rm total}), \tag{13.28}$$

where $P_{1\rm total}$ and $P_{2\rm total}$ are the total power spectra of the individual line measurements. Each of these includes terms for the power spectra of the target line, detector noise, and any contaminating lines that fall in the same band. Figure 13.8 shows the expected errors in the determination of the cross–power spectrum using the O I(63 μm) and O III(52 μm) lines at a redshift $z = 6$ for an optimized spectrometer on a 3.5-m space-borne IR telescope, providing background-limited sensitivity for 100 diffraction-limited beams covering a square on the sky 1.7° across (corresponding to 250 Mpc) and a redshift range

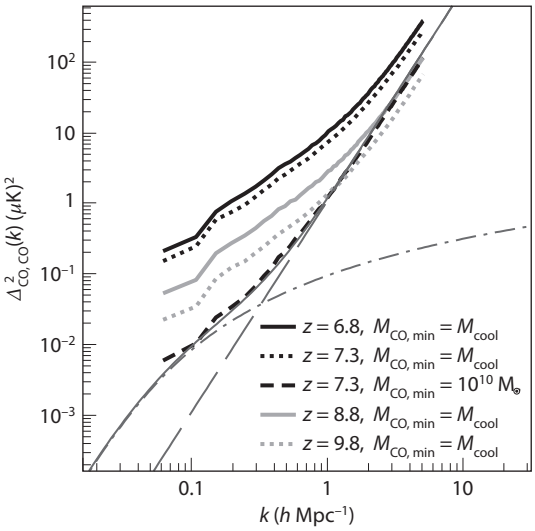


Figure 13.9 Three-dimensional power spectrum of CO(2–1) fluctuations in a simulation of reionization at several different epochs ($z = 6.8$, 7.3, 8.8, and 9.8). The CO emissivity of each galaxy is calibrated to local relations, and the total star formation rate density at $z = 6.8$ is fixed to the critical value required to maintain ionization (see text). The short-dashed line assumes that only massive galaxies emit CO radiation (possibly because the CMB dominates the dust temperature in smaller galaxies). For this scenario, we also show a simple model that includes shot noise (long-dashed curve) and an estimate of the two-halo term using the bias from the simulations (solid curve). Lidz, A. et al., *Astrophys. J.* **741**, 70 (2011). Reproduced with permission of the American Astronomical Society.

dot-dashed / AA

by / AAS. /

of $\Delta z = 0.6$ (280 Mpc) with a spectral resolution of $(\Delta\nu/\nu) = 10^{-3}$ and a total integration time of 2×10^6 seconds.

Figure 13.9 shows another example of a diffuse background in the radio, the autocorrelation of CO(2–1). (Again, this would likely be observable only if cross-correlated with another CO line, but that would only marginally affect the amplitude). The predictions are derived from a numerical simulation of reionization, and we show results for several different redshifts from the early to late stages of reionization (in the simulation, $Q_{\mathrm{H\ II}} = 0.82$ at $z = 6.8$, and 0.21 at $z = 8.8$).

The key assumptions to such a model are the mean intensity of the CO emission (which sets the overall normalization of the curves) and the luminosity–mass relationship of the source halos (which affects the shape of the curve by weighting the halos differently). The overall normalization requires two ingredients: an estimate of the total star formation rate (SFR) density and a recipe for estimating the CO luminosity as a function of SFR (and possibly other factors, like the metallicity). Here, the latter is simply calibrated to low-redshift, rapidly star-forming galaxies following equation (8.16). As described in §8.9.4,

this means that we assume that the local dust (and hence CO) excitation temperature is much larger than the CMB temperature even at these high redshifts and that the metallicity is not extremely small inside the molecular clouds.

The overall SFR density is fixed by requiring that at $z = 6.8$ it is sufficient to *keep* the Universe ionized, according to equation (9.44). Note that in the simulation, the Universe is not fully ionized at this time, but this nevertheless provides a reasonable fiducial value. At higher redshifts, the simulation assumes that the SFR is proportional to the collapse fraction $f_{\rm coll}$.

As discussed in §8.9.4, the expected brightness of the CO lines depends on the detailed physics of the ISM of high-redshift galaxies, which is currently essentially unconstrained by observations. Thus, the overall amplitude of this signal is very uncertain. The short-dashed line in Figure 13.9 illustrates this with a model in which galaxies with $M < 10^{10}\,M_\odot$ are invisible in CO (but still produce stars), possibly because their relatively small SFRs are not enough to excite the gas temperature above the CMB temperature. In this case the signal declines by roughly a factor of 1,000 on large scales, simply because of the drastically reduced overall emissivity of the gas.

Figure 13.9 reports the observed fluctuation amplitude in brightness temperature units, for which the typical value is $\sim 1\,\mu{\rm K}^2$ on moderately large scales. The shape of the power spectrum depends on both the large-scale clustering (or two-halo term in equation 13.12) and shot-noise variations in the galaxy number counts (equation 13.13), with the former dominating on large scales and the latter on small scales. The figure also shows the division between these two components for the case in which only massive galaxies emit; note that the halo-model decomposition given by these two equations provides a very good description of the signal. When less massive galaxies contribute to the CO emissivity, the shot-noise term becomes less important (because the existence of many more sources implies smaller fractional fluctuations). However, the shape of the power spectrum remains nearly the same, because the nonlinear, scale-dependent bias is important for such galaxies.

While this signal is therefore thousands of times smaller than the spin-flip background, its appearance at much higher frequencies (observed at ∼10–50 GHz rather than ∼50–200 MHz) means that the sky noise is also much, much smaller—in fact, at these frequencies the ∼10 K noise inside the detectors dominates. Moreover, these frequencies are near those already used for CMB experiments, so this technology is well developed, with both interferometers and large focal plane arrays available on the near-term horizon.

These properties enable CO mapping with much more modest instruments than those aiming to observe the spin-flip background. Equation (12.57) suggests that a single 3.5-m dish can reach a noise level of $\sim 1\,\mu$K per $10'$ pixel (its diffraction limit) in a spectral channel of a fractional width $\Delta\nu/\nu = 0.01$ at 30 GHz after an integration time of just a few days. Thus, a survey over several tens of square degrees could be accomplished relatively easily. In combination with detailed radio observations of individual galaxies from instruments like ALMA, such surveys would provide a complete census of molecular emission from the cosmic dawn. The [C II] line, at even higher frequencies,

can provide a complementary view of the lower-density neutral ISM and its cooling mechanisms.

13.3 The Cross-Correlation of Different Probes

These diffuse backgrounds can be relatively easy to measure in absolute terms but often very difficult to isolate from the many contaminants that occupy the same wave bands. One way around this problem is to cross-correlate one measurement with another; such a correlation eliminates any foreground contaminant that is not shared by both signals, greatly easing the extraction problem. We discussed in the previous section how the cross-correlation of different radio lines can help isolate them from other radio line contaminants. Here, we consider the correlation of two very different physical probes.

Crucially, such a cross-correlation can help isolate interesting physical information as well as ease the signal processing. As an example, consider the cross-correlation of a galaxy survey with the spin-flip background. If we ignore redshift-space distortions, the cross–power spectrum will be (cf. equation 12.54, recalling that the 21-cm signal is shaped by the baryon density and ionization fields)

$$\Delta^2_{21,g}(k) = T_0\langle x_{\rm H\ I}\rangle[\Delta^2_{\delta,g}(k) + \Delta^2_{x,g}(k) + \Delta^2_{x\delta,g}(k)], \tag{13.29}$$

where the subscript g refers to the fractional overdensity in the galaxy field, and x and δ denote the same for the neutral fraction and baryon density, respectively. On the relevant (large) scales, the first of these terms, $\Delta^2_{\delta,g}$, is simply proportional to the matter power spectrum multiplied by a bias factor (for the galaxy component).

Similarly, the second term is simply the cross-correlation between matter density and ionization, multiplied by the same bias factor. We studied the behavior of this term in §9.7 and found two important results. First, on large scales, it is *negative* because the ionized bubbles appear where there are many galaxies, or conversely, the gas is neutral only where there are no galaxies. Thus, the neutral gas and galaxy fields trace opposite ends of the dark matter density. However, $\Delta^2_{x,g}$ approaches zero on scales below the bubble size, because the IGM bubble is entirely ionized regardless of its small-scale density structure (at least at the level probed by the spin-flip background). Importantly, there is a relatively sharp transition to zero in this field, though that is hidden in the spin-flip background by the other terms.

However, in the cross-correlation the turnover is more apparent, because the final term in equation (13.29) cancels $\Delta^2_{\delta,g}$ on small scales.[21] To see this, note that the sum of these two terms is the Fourier transform of the quantity

$$x_{\rm H\ I}(1)\delta(1)n_g(2) = \langle x_{\rm H\ I}\rangle[1 + \delta_x(1)]\delta(1)n_g(2) \tag{13.30}$$

$$= \langle x_{\rm H\ I}\rangle\delta(1)n_g(2) + \langle x_{\rm H\ I}\rangle\delta_x(1)\delta(1)n_g(2), \tag{13.31}$$

where the labels 1 and 2 refer to the two spatial positions, and these two terms are respectively proportional to the Fourier transforms of $\Delta^2_{\delta,g}$ and $\Delta^2_{x\delta,g}$.

However, we can explicitly write the corresponding two-point function:

insert space / PE

$$\langle x_{\mathrm{H\,I}}(1)\delta(1)n_g(2)\rangle = \int dx_{\mathrm{H\,I}}(1)\, d\delta(1) dn_g(2) x_{\mathrm{H\,I}}(1)\delta(1)n_g(2) \times p[x_{\mathrm{H\,I}}(1), \delta(1)|n_g(2)]\, p[n_g(2)], \quad (13.32)$$

where we have simply expressed the correlation through a conditional probability function and noted that the mean of y is the integral of y times its overall probability distribution. On separations much smaller than a typical bubble, the two points 1 and 2 either will be within the same ionized bubble or both will be neutral. In the former case, $x_{\mathrm{H\,I}}(1) = 0$, and so the integral gets no contribution at all. In the latter case, outside ionized bubbles, there must be no galaxies, so $n_g(2) = 0$. In this simple model, the cross-correlation therefore probes *only* the cross term $\Delta^2_{x,g}$, which contains a clear feature on the scale of the typical bubble.

One way to extract this information is therefore to conduct a galaxy redshift survey in the same volume as a 21-cm survey. However, such a project would be very difficult, because the coarse resolution and huge fields of view of the radio telescopes (typically dozens or hundreds of square degrees, with several arcminutes of resolution) provide a very poor match to near-IR galaxy surveys (which, at these high redshifts, typically subtend at best several square arcminutes, but with exquisite angular resolution). A *diffuse* probe of the galaxy field therefore would provide a much better match; in particular, the diffuse CO or [C II] backgrounds are excellent candidates, because they also produce spectral fluctuations (note that broadband fluctuations, like the stellar continuum component of the near-IR background, would not correlate as well, because each wavelength comes from multiple redshifts). Figure 13.10 illustrates how such a probe would work.

The top panel of Figure 13.11 shows the resulting cross–power spectrum for the CO(2–1) line in a numerical simulation of reionization. The bottom panel shows the cross-correlation coefficient, which is defined as

$$r_{21,\mathrm{CO}}(k) = \frac{P_{21,\mathrm{CO}}(k)}{\sqrt{P_{21}(k)\,P_{\mathrm{CO}}(k)}}. \quad (13.33)$$

This quantity will be unity for perfectly correlated fields, equal to -1 for perfectly anticorrelated fields, and zero for uncorrelated fields.

The figure shows predictions for three different stages of reionization. At every stage, the cross–power spectrum is negative on large scales, and reaches near-perfect anticorrelation at sufficiently small k. But this anticorrelation gradually turns into a nearly random association ($r_{21,\mathrm{CO}} = 0$) on small scales, with the turnover scale increasing as $Q_{\mathrm{H\,II}}$ increases, reflecting the rapidly growing ionized bubbles. This kind of cross-correlation therefore offers a clear measurement of the size of the H II bubbles, which affects the spin-flip power spectrum

in a way / AA

but that is much more difficult to extract ~~from it~~ in a model-independent manner (compare Figures 12.16 and 13.11).

Another interesting cross-correlation is between the CMB and the spin-flip background. On degree scales—much larger than the size of each ionized

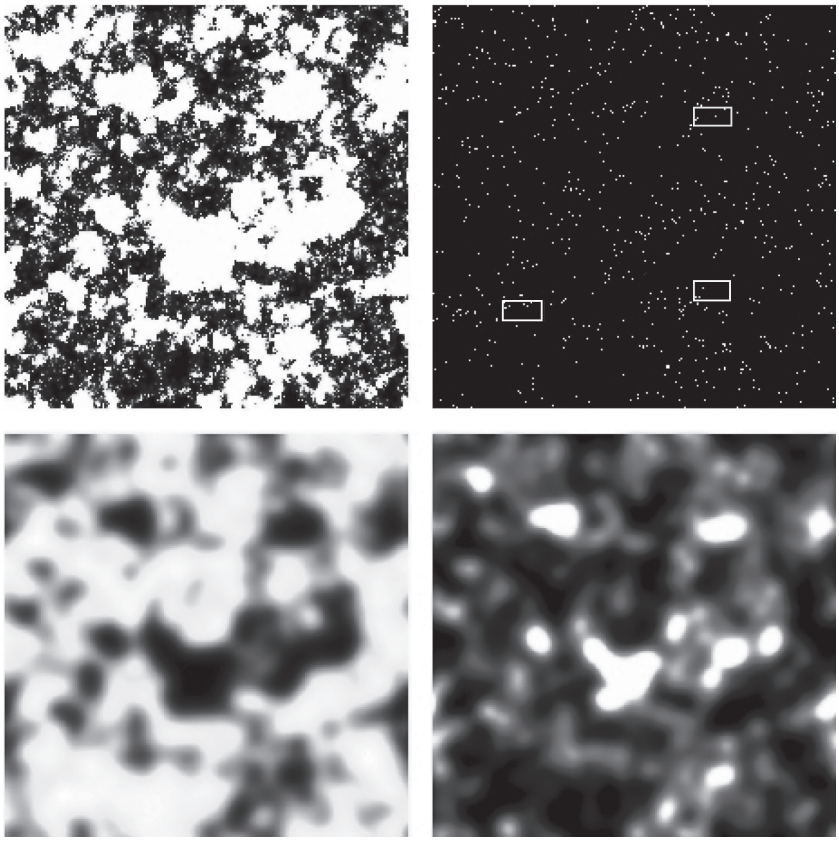


Figure 13.10 Schematic illustration of the cross-correlation between the spin-flip background and galaxy probes (see *Color Plate 32* for a color version of this figure). Each panel shows a slice from a numerical simulation ∼185 Mpc across. *Upper left:* Map of the ionized fraction midway through reionization (ionized regions are shown in white). *Upper right:* Map of the galaxy field. White squares show regions ∼ 4 × 8 arcmin across; these are each equivalent to *four* adjacent pointings of the James Webb Space Telescope. *Lower left:* Spin-flip background from this slice. *Lower right:* Galaxy map smoothed over a 6-arcmin beam, as might be observed in an intensity mapping measurement. Note the strong large-scale anticorrelation between the bottom two figures. Courtesy of / A. Lidz.

bubble—the cross-correlation is relatively easy to estimate. Qualitatively, a cross-correlation should arise because fluctuations in the density field source fluctuations in the ionized fraction (and hence 21-cm background) as well as in the baryon velocity field (which Doppler shifts CMB photons through scattering off free electrons). As we have already seen, the Doppler contribution to the CMB usually cancels out, because photons traveling across an overdense region

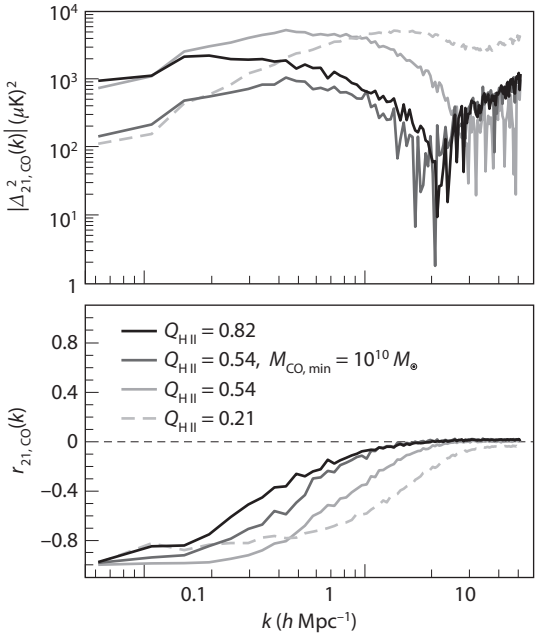


Figure 13.11 Cross-correlation between CO(2–1) emission and the spin-flip background in a numerical simulation of reionization (as in Figure 13.9). The ~~dot-dashed, dotted, and~~ solid curves take $z = 9.8$, 7.3, and 6.8 (or $Q_{H\,II} =$ 0.21, 0.54, and 0.82 in this model), ~~assuming~~ that all galaxies emit CO(2–1). The ~~long-~~dashed curve takes $z = 7.3$ but assumes that only massive galaxies emit CO. The top and bottom panels show the absolute value of the cross–power spectrum and the cross-correlation coefficient, respectively. Lidz, A., et al., *Astrophys. J.* **741**, 70 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

AA

by /
AAS. /

will be upscattered when they encounter gas falling toward the observer (on the far side of the overdensity) but will be downscattered when they encounter gas falling away from the observer. However, if the ionized fraction changes across the overdensity, the cancellation will be imperfect.[22]

If reionization were homogeneous, this would lead to an *anticorrelation* between the 21-cm brightness temperature and the CMB temperature. In this case, an overdensity would increase the 21-cm brightness but cool the CMB, as Thomson scattering would be more effective on the near side than the far side. But in the case of "inside-out" inhomogeneous reionization, we generically expect a *positive* correlation, because overdense regions host ionized bubbles (decreasing the spin-flip signal) and still cool the CMB. Unfortunately, this correlation—which in principle provides a clean probe of the evolution of the average $Q_{H\,II}$—is still quite weak, with a cross-correlation coefficient $<3\%$, because the primary CMB anisotropies dominate so strongly on the relevant scales

(multipoles $\ell \sim 100$). Only in the case that reionization occurs at very high redshifts ($z > 15$) will we be able to *detect* that there is a correlation, and even then at low confidence.[23]

On smaller angular scales, the spin-flip background and CMB should anticorrelate in the case of inhomogeneous reionization. Because the variations in the kinetic Sunyaev-Zel'dovich effect described in §13.1.2 appear where the 21-cm signal vanishes—on scales comparable to (or smaller than) the bubble size—the sign of the correlation is driven not by the velocity modes along the line of sight but by the projection $\mathbf{q}_\perp$ of the product of the velocity and electron density (see equation 13.4 and Figure 13.4). In principle, this small-scale correlation provides a clean probe of the ionized bubble properties. Unfortunately, the cross-correlation is again dwarfed by the primary CMB anisotropies: although the kinetic Sunyaev-Zel'dovich signal peaks at quite large angular scales, the cross-correlation component is confined to $\ell < 8{,}000$ (where the primary CMB signal is still large), largely because of cancellation of the structures in the integrated CMB map. Thus, although the CMB temperature-21 cm cross-correlation contains interesting physical information, it does not appear to be a useful observable in practice.[24]

13.4 The Fossil Record of the Local Group

The focus of this book is on studies of the distant Universe: direct observations of galaxies (and the objects in them) or indirect probes of their environment. But, of course, these early generations of galaxies are the progenitors of today's galaxies, including the Milky Way, which must therefore contain remnants and signatures of these first structures. Here, we briefly discuss the prospects for *stellar archaeology* and its utility in understanding the cosmic dawn.

The hierarchical structure formation paradigm implies that the small dark matter halos, in which the first stars and galaxies formed, merged to form increasingly larger galaxies over time. During these violent merger events, existing gas reservoirs in galaxy cores would likely have undergone compression and formed stars, and continuing accretion at late times would have formed even more stars. Thus, the large majority of stars in today's galaxies formed long after the cosmic dawn.

But what of the *existing* stars within the merging halos, that may have formed in pristine conditions (Population III) or shortly after, with very low metallicity? Recent large multiobject spectroscopic surveys have uncovered several hundred *extremely metal-poor* stars, characterized as having a relative abundance of iron to hydrogen at least three orders of magnitude smaller than that in the Sun (denoted as $[\mathrm{Fe/H}] < -3$). Stellar archaeologists hope to use these stars to uncover information like the IMF and efficiency of second-generation star formation and the nucleosynthetic yields of the very first stars, which may have provided the heavy elements for the extremely metal-poor stars.

A related question is whether the small galaxies that surround the Milky Way—many containing so few stars that they remained hidden until the most

recent generation of large surveys—can be traced back to the cosmic dawn. If so, they may carry imprints of important feedback mechanisms, such as initial metal enrichment, the growth of the Lyman-Werner background, or reionization. For example, galaxies with halo masses $M_h \sim 10^8\, M_\odot$ may have formed some stars before reionization and then been shut off by the photoheating that accompanied that event (see §9.9.2). Because these systems would then have undergone relatively little star formation in the past 10 gigayears, they may carry more obvious records of these events than large galaxies like the Milky Way.

13.4.1 Stellar Archaeology

To successfully use metal-poor stars to understand the earliest generations of structure requires a number of inputs. First, one must *find* such stars, which is no easy task and is becoming possible only with large spectroscopic surveys that can identify very rare objects. Second, one must understand the dynamics of these stars and where they may have originated. Finally, one must relate their chemical abundances to the stars that enriched them, which requires detailed models of massive star supernovae, metal mixing in the ISM, and second-generation star formation.

- *Where are the extremely metal-poor stars?* So far, observations have found most of these stars in the outer halo of our Milky Way, which appears to have a somewhat lower mean metallicity (centered at [Fe/H] ~ -2.2) than the inner halo (by a factor of 3).[25]

  Numerical simulations are consistent with this picture, showing that extremely metal-poor stars should appear throughout the Galaxy. This is largely because metal enrichment throughout the IGM is highly inhomogeneous, so that pockets of metal-free gas may have persisted until relatively late in the Universe's history, from $z \sim 5$–3. These relics would then have been incorporated into the outer halo of the galaxy.[26]
- *Where are the* oldest *extremely metal-poor stars?* If galaxies were composed solely of dark matter and stars, numerical simulations of hierarchical structure formation models would provide a fairly robust answer to this question: near the centers of galaxies. These simulations show that galaxies form "inside-out," with the first objects to be accreted (i.e., the most overdense nearby regions, where the first stars would also have formed) residing closest to the bottom of the galaxy's potential well, and later additions located increasingly farther out in the galaxy's halo. Thus, although extremely metal-poor stars may be spread throughout the halo, the *oldest* will be located near the center.[27] This presents significant problems for surveys, as these few old stars would be buried in the much more numerous stars of our Galactic bulge and be subject to relatively large extinction.

  However, baryonic processes may mitigate this difficulty to some extent. In particular, spiral perturbations driven by accretion events can cause stars to migrate over large radial distances and get deflected to much larger orbits.[28] If so, such stars may be much more accessible to searches in the

outskirts of galaxies, although their spatial distribution after a sequence of such events has not been well quantified.

- *What are the "chemical fingerprints" of the first stars?* Once a set of such stars is found, stellar archaeologists hope to use the stars, abundance patterns and other properties to learn about star formation in the early Universe—both at the time these objects formed and in the earlier generations that enriched their fuel supply. The simplest approach is to use the "chemical fingerprints" present in these stars' abundance patterns to deduce the properties of the precursor stellar populations' supernovae.

  Such efforts have a long tradition in astronomy, dating to efforts to understand abundance patterns within our own solar system and in nearby stars. However, although the general problem is well posed, extracting quantitative information remains difficult. Astrophysicists have a good qualitative understanding of the nuclear pathways through which heavy elements form. Broadly, there are two different processes: In the *r-process*, which occurs during supernovae, neutrons are added to seed nuclei (usually ^{56}Ni) much more rapidly than β decay can occur. The resulting nuclei form a distinct pattern set by the locations of closed neutron shells, where the cross section for continued neutron capture drops rapidly. However, the shells that form during the r-process are overabundant in neutrons and so suffer a sequence of β decays before they reach stability. The contrasting *s-process* occurs when neutrons are added over long timescales, so that the nuclei can undergo β decay and grow through a sequence of stable nuclei. This process occurs largely in the atmospheres of asymptotic giant branch stars, over longer timescales than supernova nucleosynthesis.

  Interestingly, the chemical patterns produced in supernovae depend strongly on the properties of their progenitors.[29] As discussed in §5.5, stars of $\sim$140–260 $M_\odot$ are subject to a pair-production instability, in which much of the star's internal energy is lost when photon collisions create electron–positron pairs, locking up much of the thermal energy in the rest mass of those particles. The star then explodes violently. Before the explosion of pair-instability supernovae, the star has only a very small excess of neutrons, which strongly suppresses the formation of elements with an odd atomic number as compared with an even one. Moreover, these stars have large oxygen-fusing regions, which leads to an overproduction of silicon and sulfur compared with the composition of more normal supernovae.

  Assuming that the extremely metal-poor stars were formed from gas enriched by only one or a few supernovae in the early generation of stars, one might therefore expect to see such fingerprints in their abundance patterns. However, to date this has proved not to be the case: instead, these stars—just like most of those in the Milky Way—appear to have been enriched by supernovae from stars with masses $\sim$10–40 $M_\odot$, based largely on their overabundance of so-called α elements. These are nuclei made up of integer multiples of helium nuclei and are synthesized in the silicon-burning phase before such stars explode. Thus, there is so far little evidence for earlier generations of very massive ($>$100 $M_\odot$) stars in these

Galactic searches, although there are extragalactic clues for possible pair-instability supernova explosions[30] and their odd–even abundance pattern in damped Lyman-α systems.[31].

There are also several interesting anomalies that currently remain to be understood in the abundances of particular elements within the extremely metal-poor population. One interesting feature is the large scatter in carbon and nitrogen, which can be greatly enhanced relative to iron. Many (but not all) of these stars also have enhanced s-process abundances; most likely, they were therefore polluted by an intermediate-mass binary companion, which makes it difficult to tease out the abundances of the precursor supernovae.

The cooling rate by atomic carbon and oxygen dominates that of molecular hydrogen once their abundances exceed $\sim 10^{-3.5}$ of the solar values. If such enhanced cooling is a prerequisite for the ability of the gas to fragment into low-mass stars, then one would expect all low-mass stars in the Milky Way halo to show carbon or oxygen abundances above this threshold (see §6.5). Figure 13.12 shows that existing data are consistent with this theoretical expectation.

- *What can we learn about the IMF of the earliest stars?* The mere fact that no *metal-free* stars have been found—despite the relatively large number of extremely metal-poor stars—suggests that the very first generation of stars was skewed toward high masses, with no evidence for Population III stars with a mass $<0.8\ M_\odot$. The relatively common carbon enhancement found in these stars also points to a higher characteristic mass, since a large fraction of the binary companions would have had to be of relatively large mass as well to evolve so as to donate some of their material to the observed stars.

  However, the lack of clear signatures of pair-instability supernovae, with masses $>100\ M_\odot$, and the relative success of "normal" supernovae at reproducing the abundance patterns of extremely metal-poor stars, argues against very massive stars having been common even in the earliest phases of structure formation—in contrast with the results of most numerical simulations. However, such an interpretation assumes that the heavy-element products of the supernova mix efficiently with the ISM of the galaxy; if, instead, the mixing occurs only slowly, the second-generation stars may actually have relatively high metallicities and so lie outside the target area of existing surveys. A great deal remains to be learned from these surveys, and improved modeling of the transition from one stellar population to the next will be a crucial element of extracting the best possible information.
- *What additional information can we extract from these surveys?* It has become increasingly clear that in the local Universe, star formation generally proceeds inside massive star clusters of uniform abundance (albeit of widely varying masses). If similar processes occurred during the early generations of star formation, we would expect to find "clumps" in the abundance patterns characteristic of these star clusters, even if the stars themselves have

should be apostrophe, not comma / PE

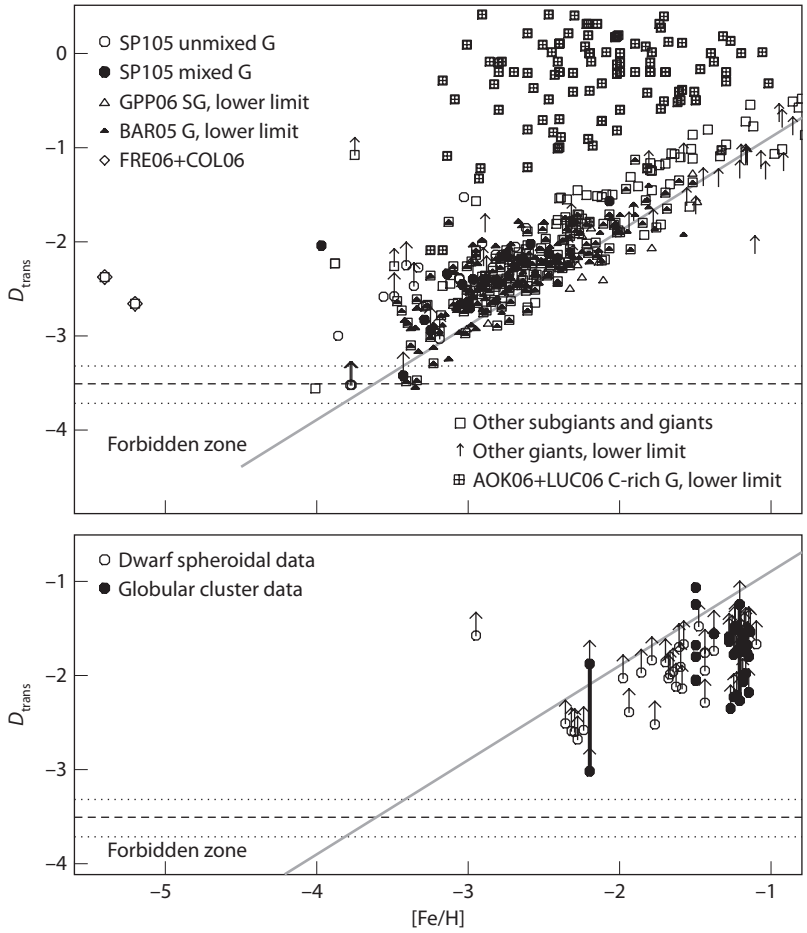


Figure 13.12 Measure of the carbon and oxygen abundance for metal-poor stars, $D_{\text{trans}} \equiv \log\left(10^{[C/H]} + 0.3 \times 10^{[O/H]}\right)$, as a function of the iron abundance relative to hydrogen, [Fe/H]. *Top panel*: Galactic halo stars. *Bottom panel*: Stars in dwarf spheroidal galaxies and globular clusters. G and SG indicate giants and subgiants. The critical limit marked with a dashed line was predicted theoretically by comparing the cooling rate for carbon and oxygen lines with H_2 cooling (which controls the formation of massive Population III stars). Without metal-line cooling, no low-mass stars should be found below this line. The dotted lines define the uncertainty in the theoretical prediction without dust cooling. If dust cooling is added, the dashed line is lowered by two orders of magnitude. Interestingly, all data points are well above the theoretical lines for metal and dust cooling so far. Frebel, A., Johnson, J. L., & Bromm, V., *Mon. Not. R. Astron. Soc.* **380**, L49 (2007). Copyright 2007 by the Royal Astronomical Society.

dispersed on widely varying orbits. Such clumps could tell us about the environments in which early stars formed and the processes that regulated their birth; they would also provide "uniform" samples of stars over which we could average to study metal abundance across larger scales than individual stars, and with which we could account for anomalies like mass transfer between binary companions.

13.4.2 Ultrafaint Dwarf Galaxies around the Milky Way

Recent large spectroscopic surveys have revealed a wealth of information about the structure of our Milky Way galaxy and its immediate neighbors. In particular, they have enabled the discovery of a sizable class of "ultrafaint" dwarf galaxies, with a total luminosity $<10^5\ L_\odot$. These are particularly intriguing objects, because they appear to have undergone only one or a few star formation events (rather than the rich history common to larger galaxies), so the current stellar populations offer relatively clean tracers of the first generations of stars in these particular dark matter halos. Moreover, these galaxies are small enough that one can realistically imagine obtaining a fairly complete census of their stars.

The present data show that these dwarf galaxies have low average metallicities, some as small as [Fe/H] ~ -2.6, and substantial scatter in the abundances (at least an order of magnitude). Overall, the abundance patterns resemble those in the extremely metal-poor Galactic halo stars described in §13.4.1.[32]

Thus, it seems that one of the feedback mechanisms that we described earlier was responsible for shutting down ongoing star formation in these galaxies. Three possibilities immediately come to mind: *(1)* UV feedback from the first stars that photodissociated H_2 and terminated cooling in minihalos below the atomic cooling threshold; *(2)* the supernova feedback generated by the first wave of star formation in the dwarf galaxy itself; and *(3)* photoheating from reionization (or an even earlier stage) that suppressed accretion onto these small halos, or possibly even evaporated any existing gas.

The canonical theoretical picture assumes that the initial burst of star formation in an isolated region contained only very high mass Population III stars. If the ultrafaint dwarfs—which contain low-mass stars—were then to form in the same dark matter halos, those galaxies must have been able to retain their gas (or at least reaccrete it) after these stars died. This process generally argues against extremely low mass halos with shallow gravitational potential wells for the dwarfs. It also poses a challenge for any models that attribute the end of star formation in these galaxies to supernova winds: although such winds could have arisen from a later generation of star formation, the binding energy of the halo scales as $\propto M_h^2$, so one naively expects halos to become more stable as they grow.

The Lyman-Werner background on its own is also unlikely to be the factor stopping star formation in these galaxies. It is important only when H_2 dominates the cooling function; thus, because the stellar populations in these dwarfs have nonzero metallicity, it must not have been the dominant coolant in these

populations (even if it was responsible for slowing down or stopping star formation in earlier phases).

The other key question is then how the first and second generations of star formation are related spatially. If minihalos could either retain or reacquire their gas after the initial burst of Population III star formation, they may have been able to form low-mass Population II stars that survive to the present day. If, however, these halos instead lost their gas for a long period of time, larger objects—possibly above the atomic cooling threshold—would have held the bulk of the second generation of stars. Moreover, because in that case these new stars formed in a different halo than their Population III progenitors, it is easier to imagine that they formed in a burst mode that was able to evacuate the galaxy of its remaining gas, shutting down later star formation. Thus, understanding the dynamics and contents of these ultrafaint dwarfs offers the tantalizing possibility of constraining the large-scale pathways that enabled high-redshift star formation.

More detailed information is also available from these objects: for example, if an initial starburst did clear out the dwarf's gas, the stellar population should have little evidence of "self-enrichment." Rather, the stars might all have abundances characteristic of core collapse (and possibly Population III) supernovae, without any substantial s-process elements. The observed scatter in the metallicity within individual dwarfs also suggests that metals must not have been efficiently mixed across galactic scales, at least if the picture of a single burst of star formation is correct. Interestingly, old globular clusters have little or no apparent scatter in their abundances, which implies much more efficient mixing in such systems.

The recent discovery of these dwarfs, and the rapidly increasing samples of metal-poor stars inside them and inside our own Galaxy, have opened a new window into studies of the impact of star formation in the early Universe. The implications of these studies are now only beginning to be understood, and a great deal of work on both the observational and theoretical ends is needed to disentangle the clues lying within. Stellar archaeology promises to remain a fascinating frontier for many years to come.

Appendix A

Useful Numbers

Fundamental Constants	
Newton's constant (G)	$= 6.67 \times 10^{-8}$ cm^3 g^{-1} s^{-2}
Speed of light (c)	$= 3.00 \times 10^{10}$ cm s^{-1}
Planck's constant (h)	$= 6.63 \times 10^{-27}$ erg s
Electron mass (m_e)	$= 9.11 \times 10^{-28}$ g $\equiv 511$ keV$/c^2$
Electron charge (e)	$= 4.80 \times 10^{-10}$ esu
Proton mass (m_p)	$= 1.67 \times 10^{-24}$ g $= 938.3$ MeV$/c^2$
Boltzmann's constant (k_B)	$= 1.38 \times 10^{-16}$ erg K^{-1}
Stefan-Boltzmann constant ($\sigma_{\rm SB}$)	$= 5.67 \times 10^{-5}$ erg cm^{-2} s^{-1} K^{-4}
Radiation constant ($a_{\rm rad}$)	$= 7.56 \times 10^{-15}$ erg cm^{-3} K^{-4}
Thomson cross section (σ_T)	$= 6.65 \times 10^{-25}$ cm^2
Astrophysical Numbers	
Solar mass ($M_\odot$)	$= 1.99 \times 10^{33}$ g
Solar radius ($R_\odot$)	$= 6.96 \times 10^{10}$ cm
Solar luminosity ($L_\odot$)	$= 3.9 \times 10^{33}$ 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)	$= 3.086 \times 10^{18}$ cm
1 kiloparsec (kpc)	$= 10^3$ pc
1 megaparsec (Mpc)	$= 10^6$ pc
1 gigaparsec (Gpc)	$= 10^9$ pc
1 astronomical unit (AU)	$= 1.5 \times 10^{13}$ cm
1 year (yr)	$= 3.16 \times 10^7$ s
1 light year (ly)	$= 9.46 \times 10^{17}$ cm
1 eV	$= 1.60 \times 10^{-12}$ ergs $\equiv 11,604$ K $\times\ k_B$
1 erg	$= 10^{-7}$ J
Photon wavelength ($\lambda = c/\nu$)	$= 1.24 \times 10^{-4}$ cm (photon energy/1 eV)$^{-1}$
1 nanoJansky (nJy)	$= 10^{-32}$ erg cm^{-2} s^{-1} Hz^{-1}
1 Angstrom (Å)	$= 10^{-8}$cm

change to `=' /AA

insert space (before cm) / PE

Unit Conversions	
1 micron (μm)	$= 10^{-4}$cm
1 km s^{-1}	= 1.02 pc per million years
1 arcsecond ($''$)	$= 4.85 \times 10^{-6}$ radians
1 arcminute ($'$)	$= 60''$
1 degree ($^\circ$)	$= 3.6 \times 10^{3}''$
1 radian	$= 57.3^\circ$

insert space (before cm) / PE

Appendix B

Cosmological Parameters

We list here the cosmological parameters assumed throughout the text. These choices are based on Komatsu, E., et al. *Astrophys. J. Suppl.* **180**, 330 (2009).

Cosmological Parameters			
Matter density	Ω_m	$=$	0.28
Baryon density	Ω_b	$=$	0.05
Dark energy density	Ω_Λ	$=$	0.72
Dark energy equation of state	w	$=$	-1
Hubble constant	h	$=$	0.7
Scalar index	n_s	$=$	1
Power spectrum normalization	σ_8	$=$	0.82
Helium mass fraction	Y_p	$=$	0.24

Notes

Chapter 1

[1] Pen, U.L. *Astron. Astrophys. Suppl. Seri.* **120**, 49 (1996).
[2] Loeb, A. *Phys. Rev.* **D** 65, 7301 (2002); Nagamine, K., & Loeb, A., *New Astronomy* **8**, 439 (2003) and **9**, 573 (2004); Cox, T. J., & Loeb, A., *Mon. Not. R. Astron. Soc.* **386**, 461 (2008).
[3] Loeb, A. *J. Cosm. & Astroparticle Phys.* **4**, 23 (2011).

Chapter 2

[1] An analytic expression for the growth factor in terms of special functions was derived by D. Eisenstein at http://arxiv.org/pdf/astro-ph/9709054v2.
[2] Lahav, O., et al. *MNRAS*, **251**, 128 (1991).
[3] Peebles, P.J.E. *Principles of Physical Cosmology*. Princeton University Press (1993), p. 626.
[4] Eisenstein,D. J., & Hu, W. *Astrophys. J.* **511**, 5 (1999); Padmanabhan, T., *Theoretical Astrophysics,* Vol. III*: Galaxies and Cosmology*, Cambridge University Press. (2002), pp. 319–320.
[5] Loeb, A., & Zaldarriaga, M. *Phys. Rev.* **D71**, 3520 (2005).
[6] Tseliakhovich, D., & Hirata, C. *Phys. Rev.* **D82**, 3520 (2010).

Chapter 3

[1] See, e.g., Bertschinger, E. *Astrophys. J. Suppl.* **58**, 39 (1985); and Fillmore, J. A. & Goldreich, P. *Astrophys. J.* **281**, 1 (1984).
[2] Loeb, A., & Zaldarriaga, M. *Phys. Rev.* **D71**, 103520 (2005).
[3] Gnedin, N.Y., & Hui, L. *Mon. Not. R. Astron. Soc.* **296**, 44 (1998). Note that we associate the filtering radius with π/k_F in analogy with the Jeans radius $\pi/k_{\rm J}$; hence, our filtering mass M_F is one-eighth that defined by Gnedin & Hui, who adopt a filtering radius of $2\pi/k_F$.
[4] Bryan, G., & Norman, M. *Astrophys. J.* **495**, 80 (1998).
[5] Navarro, J. F., Frenk, C. S., & White, S.D.M. *Astrophys. J.* **490**, 493 (1997).
[6] Navarro, J. F., et al. *Mon. Not. R. Astron. Soc.* **402**, 21 (2010).
[7] Press, W. H., & Schechter, P. *Astrophys. J.* **187**, 425 (1974).
[8] Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. *Astrophys. J.* **379**, 440 (1991).
[9] Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. *Astrophys. J.* **379**, 440 (1991); Lacey, C. G., & Cole, S. *Mon. Not. R. Astron. Soc.* **262**, 627 (1993).

[10] Lacey, C. G., & Cole, S. *Mon. Not. R. Astron. Soc.* **262**, 627 (1993).
[11] Sheth, R. K., & Tormen, G. *Mon. Not. R. Astron. Soc.* **329**, 61 (2002).
[12] Sheth, R., & Tormen, G., *Mon. Not. R. Astron. Soc.* **329**, 61 (2002)
[13] Reed, D., et al. *Mon. Not. R. Astron. Soc.* **374**, 2 (2007)
[14] Tinker, J.,et al. *Astrophys. J.* **688**, 709 (2008).
[15] Mo, H. J., & White, S.D.M., *Mon. Not. R. Astron. Soc.* **282**, 347 (2006).
[16] Scherrer, R. J., & Bertschinger, E. *Astrophys. J.* **381**, 349 (1991).
[17] Kravtsov, A. V., et al. *Astrophys. J.* **609**, 35 (2004).

Chapter 4

[1] Zeldovich, Ya. B. *Astron. Astrophys.* **5**, 84 (1970).
[2] Shandarin, S. F., & Zel'dovich,Ya. B. *Rev. Mod. Phys.* **61**, 185 (1989).
[3] Scheuer, P.A.G. *Nature* **207**, 963 (1965).
[4] Gunn, J. E., & Peterson, B. A. *Astrophys. J.* **142**, 1633 (1965).
[5] Osterbrock, D. E. *Astrophysics of Gaseous Nebulaeand Active Galactic Nuclei*. Mill Valley, CA: University Science Books (1989).
[6] Verner, D. A., Ferland, G. J., Korista, T., & Yakovlev, D. G. *Astrophys. J.* **465**, 487 (1996).
[7] Osterbrock, D. E. *Astrophysics of Gaseous Nebulaeand Active Galactic Nuclei*. Mill Valley, CA: University Science Books (1989).
[8] Fan, X., et al. *Astron. J.* **132**, 117 (2006).
[9] Songaila, A., & Cowie, L. L. *Astrophys. J.* **721**, 1448 (2010).
[10] Mesinger, A., & Furlanetto, S. R., *Mon. Not. R. Astron. Soc.* **400**, 1461 (2009).
[11] Schaye, J., et al. *Astrophys. J.* **596**, 768 (2003); Aguirre, A., et al. *Astrophys. J.*, **602**, 38 (2004).
[12] Fan, X., et al. *Astron. J.* **132**, 117 (2006); McGreer, I. D., Mesinger, A., & Fan, X. *Mon. Not. R. Astron. Soc.* **415**, 3237 (2011).
[13] Oh, S. P., & Furlanetto, S. R. *Astrophys. J.* **620**, L9(2005).

Chapter 5

[1] Haiman, Z., Thoul, A. A., & Loeb, A. *Astrophys. J.* **464**, 523 (1996); Tegmark, M., et al. *Astrophys. J.* **474**, 1 (1997).
[2] Tegmark, M., et al. *Astrophys. J.* **474**, 1 (1997).
[3] Hirata, C. M., & Padmanabhan, N. *Mon. Not. R. Astron. Soc.* **372**, 1175 (2006).
[4] Abel, T., Bryan, G. L., & Norman, M. L. *Science* **295**, 93 (2002); Bromm, V., et al. *Astrophys. J.* **564**, 23 (2002).
[5] Tegmark, M., et al. *Astrophys. J.* **474**, 1 (1997).
[6] Tegmark, M., et al. *Astrophys. J.* **474**, 1 (1997).
[7] Abel, T., Bryan, G. L., & Norman, M. L., *Science* **295**, 93 (2002);Bromm, V., et al. *Astrophys. J.* **564**, 23 (2002).
[8] These reactions rates are highly uncertain; we take fiducial values from Palla, F., Salpeter, E. E., & Stahler, S. W., *Astrophys. J.* **271**, 632 (1983).
[9] Tan., J. C., & McKee, C. F. *Astrophys. J.* **603**, 383 (2004).

[10] Tan., J. C., & McKee, C. F. *Astrophys. J.* **603**, 383 (2004).
[11] The numerical estimates of the efficiencies of the following four feedback mechanisms are taken from McKee, C. F., & Tan, J. C. *Astrophys. J.* **681**, 771 (2008).
[12] Clark, P. C., et al. *Science* **331**, 1040 (2011).
[13] Clark, P. C., et al. *Astrophys. J.* **727**, 110 (2011).
[14] Pudritz, R. E. *Science* **295**, 68 (2002), and references therein.
[15] Salpeter, E. *Astrophys. J.* **121**, 161 (1955).
[16] Rees, M. J. *Mon. Not. R. Astron. Soc.* **176**, 483 (1976).
[17] Ebert, R. *Zeitschrift für Astrophysik* **37**, 217 (1955); Bonnor, W. B. *Mon. Not. R. Astron. Soc.* **116**, 351 (1956).
[18] Oh, S. P., & Haiman, Z. *Astrophys. J.* **569**, 558 (2002).
[19] Johnson, J. L., & Bromm, V. *Mon. Not. R. Astron. Soc.* **366**, 247 (2006).
[20] Yoshida, N., et al. *Astrophys. J.* **663**, 687 (2007).
[21] Bromm, V., Kudritzki, R. P., & Loeb, A. *Astrophys. J.* **552**, 464 (2001).
[22] Tumlinson, J., & Shull, M. J. *Astrophys. J.* **528**, L65 (2000).
[23] Heger, A., et al. *Astrophys. J.* **591**, 288 (2003).
[24] Heger, A., & Woosley, S. E. *Astrophys. J.* **567**, 532 (2002).
[25] Klebesadel, R. W., Strong, I. B., & Olson, R. A. *Astrophys. J.* **182**, L85 (1973).
[26] Zhang, W., Woosley, S. E., & MacFadyen, A. I. *J. of Phys. Conf. Ser.* **46**, 403 (2006).
[27] Cucchiara, A., et al. *Astrophys. J.* **736**, 7 (2011).

Chapter 6

[1] See Stecher, T. P., & Williams, D. A. *Astrophys. J.* **149**, L29 (1967) for a quantitative analysis.
[2] Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **476**, 458 (1997); **484**, 985 (1997).
[3] Draine, B. T., & Bertoldi, F. *Astrophys. J.* **468**, 269 (1996).
[4] Glover, S.C.O., & Brand, P.W.J.L. *Mon. Not. R. Astron. Soc.* **321**, 385 (2001).
[5] Haiman, Z., Rees, M. J., & Loeb, A. *Astrophys. J.* **467**, 522 (1996).
[6] Oh, S. P. *Astrophys. J.* **553**, 449 (2001); Venkatesan, A., Giroux, M. L., & Shull, J. M. *Astrophys. J.* **563**, 1.
[7] Machacek, M. E., Bryan, G. L., & Abel, T. *Mon. Not. R. Astron. Soc.* **338**, 273 (2003).
[8] This and the following expressions are derived in detail in Franco, J., Tenorio-Tagle, G., & Bodenheimer, P. *Astrophys. J.* **349**, 126 (1990).
[9] Whalen, D., Abel, T., & Norman, M. L. *Astrophys. J.* **610**, 14 (2004).
[10] Whalen, D., et al. *Astrophys. J.* **610**, 14 (2004).
[11] Yoshida, N. et al. *Astrophys. J.* **663**, 687 (2007).
[12] Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **391**, 457 (2008); **396**, 377 (2009).
[13] Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **391**, 457 (2008).
[14] Murray, N., Quataert, E., & Thompson, T. A., *Astrophys. J.* **618**, 569 (2005).

[15] Thompson, T. A., Quataert, E., & Murray, N. *Astrophys. J.* **630**, 167 (2005).
[16] Mac Low, M.-M., & Ferrara, A. *Astrophys. J.* **513**, 142 (1999).
[17] Bromm, V., & Loeb, A. *Nature* **425**, 812 (2003).
[18] Schneider, R., et al. *Mon. Not. R. Astron. Soc.* **369**, 1437 (2006).
[19] Voit, G. M. *Astrophys. J.* **465**, 548 (1996).
[20] Genzel, R., & Karas, V. *IAU Symp.* **238**, 173 (2007); Ghez, A., et al. *Astrophys. J.* **689**, 1044 (2008).

Chapter 7

[1] Begelman, M. C., Blandford, R. D., & Rees, M. J. *Rev. Mod. Phys.* **56**, 255 (1984).
[2] Wyithe, J.S.B., & Loeb, A. *Astrophys. J.* **595**, 614 (2003).
[3] Bardeen, J. M., Press, W. H., & Teukolsky, S. A. *Astrophys. J* **178**, 347 (1972); or see chapter 14 in Shapiro, S. L., & Teukolsky, S. A. *Black Holes, White Dwarfs, and Neutron Stars*. New York: Wiley (1982).
[4] See Shapiro, S. L., & Teukolsky, S. A. *BlackHoles, White Dwarfs and Neutron Stars: The Physics of Compact Objects*. New York: Wiley (1983), chapter 14.
[5] Shakura, N. I., & Sunyaev, R. A. *Astron. & Astrophys.* **24**, 337 (1973); Novikov, I. D., & Thorne, K. S. in *Black Holes ed.* C. DeWitt and B. DeWitt. New York: Gordon and Breach (1973).
[6] 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).
[7] Haiman, Z., Kocsis, B., & Menou, K. *Astrophys. J.* **700**, 1952 (2009).
[8] Milosavljević, M., & Loeb, A. *Astrophys. J.* **604**, L45 (2004).
[9] Narayan, R., & McClintock, J. *New Astronomy Reviews* **51**, 733 (2008), and references therein.
[10] Loeb, A. *Phys. Rev.* **D37**, 3484 (1988).
[11] See Loeb, A., *Phys. Rev.* **D81**, 7503 (2010), and references therein.
[12] 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).
[13] Blecha, L., & Loeb, A. *Mon. Not. R. Astron. Soc.* **390**, 1311 (2008); Tanaka, T., & Haiman, Z. *Astrophys. J.* **696**, 1798 (2009).
[14] Shapiro, S. L., & Teukolsky, S. A. *Black Holes, White Dwarfs and Neutron Stars: The Physics of Compact Objects*. New York: Wiley (1983), chapter 17.
[15] Gültekin, K., et al. *Astrophys. J.* **698**, 198 (2009).
[16] H. Netzer, & Laor A., *Astrophys. J.* **404**, L51, (1993).
[17] Bentz, M., et al. *Astrophys. J.* **705**, 199 (2009).
[18] Wang, R., et al. *Astrophys. J.* **714**, 699 (2010).
[19] See, e.g., Wyithe, J.S.B., & Loeb, A. *Astrophys. J.* **595**, 614 (2003); Hopkins, P. F., & Hernquist, L. *Astrophys. J.* **698**, 1550 (2009).
[20] Dietrich, M., & Hamann, F. *Rev. Mex. de Astron. Astrof. Conf. Ser.* **32**, 65 (2008).
[21] A. Laor, & Draine B., *Astrophys. J.* **402**, 441, (1993).

[22] Faucher-Giguere, C.-A., Quataert, E., & Murray, N. *Mon. Not. R. Astron. Soc.*, submitted (arXiv.org/1108.0413) (2011).
[23] M. Colpi, Mayer L., & Governato F., *Astrophys. J.* **525**, 720 (1999).
[24] Binney, J., & Tremaine, S. *Galactic Dynamics*, Princeton, NJ: Princeton University Press (2008).
[25] Peters, P. C. *Phys. Rev.* **136**, 1224 (1964).
[26] Hoffman, L., & Loeb, A. *Mon. Not. R. Astron. Soc.* **377**, 957 (2007).
[27] O'Leary, R. M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **395**, 781 (2009); and http://arxiv.org/pdf/1102.3695v2.pdf.
[28] Peters, P. C. *Phys. Rev.* **136**, 1224 (1964).
[29] See Shapiro, S. L., & Teukolsky, S. A. *Black Holes,White Dwarfs and Neutron Stars: The Physics of Compact Objects*. New York: Wiley (1983), pp. 469–472.
[30] Holz, D., & Hughes, S. *Astrophys. J.* **629**, 15 (2005).
[31] http://lisa.nasa.gov/.
[32] http://www.advancedligo.mit.edu/.
[33] See, e.g., Kocsis, B., & Loeb, A. *Phys. Rev. Lett.* **101**, 041101 (2008); Lippai, Z., Frei, Z., & Haiman, Z. *Astrophys. J. Lett.* **676**, L5 (2008); Schnittman, J. D., & Krolik, J. H. *Astrophys. J.* **684**, 835 (2008).
[34] See, e.g., Loeb, A. *Phys. Rev. Lett.* **99**, 041103 (2007); Blecha, L., Cox, T. J., Loeb, A., & Hernquist, L. *Mon. Not. R. Astron. Soc.* **412**, 2154 (2011).
[35] Stone, N., & Loeb, A. *Mon. Not. R. Astron. Soc.* **412**, 75 (2011).
[36] The first detection of a relativistic jet from the tidal disruption of a star by a supermassive black hole at a cosmological distance was reported by Burrows D. N., et al. *Nature* **476**, 421 (2011); Bloom, J. S., et al. *Science* **333**, 203 (2011); and Zauderer, A., et al. *Nature* **476**, 425 (2011).

Chapter 8

[1] Birnboim, Y., & Dekel, A. *Mon. Not. R. Astron. Soc.* **345**, 349 (2003).
[2] Neistein, E., van den Bosch, F. C., & Dekel, A. *Mon. Not. R. Astron. Soc.* **372**, 933 (2006).
[3] Cox, T. J., et al. *Mon. Not. R. Astron. Soc.* **384**, 386 (2008).
[4] Mo, H. J., Mao, S., & White, S.D.M. *Mon. Not. R. Astron. Soc.* **295**, 319 (1998).
[5] Wood, K., & Loeb, A. *Astrophys. J.* **545**, 86 (2000).
[6] Barkana, R., & Loeb, A. *Astrophys. J.* **531**, 613 (2000).
[7] Warren, M. S., Quinn, P. J., Salmon, J. K., & Zurek, W. H. *Astrophys. J.* **399**, 405 (1992).
[8] 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).
[9] Krumholz, M. R., & McKee, C. F. *Astrophys. J.* **630**, 250 (2005); Krumholz, M. R., & Tan, J. C. *Astrophys. J.* **654**, 304 (2007).
[10] Ibid.
[11] Robertson, B. E., & Kravtsov, A. V. *Astrophys. J.* **680**, 1083 (2008).
[12] Thompson, T. A., Quataert, E., & Murray, N. *Astrophys. J.* **630**, 167 (2005).
[13] Ibid.

[14] See Hopkins, P. F., & Quataert, E. *Mon. Not. R. Astron. Soc.* **415**, 1027 (2011), and references therein.
[15] Kulkarni, G., & Loeb, A. *Mon. Not. R. Astron. Soc.*, submitted, arXiv.org/1107.0517 (2011).
[16] Dekel, A., & Woo, J. *Mon. Not. R. Astron. Soc.* **344**, 1131 (2003); Kauffmann, G., et al. *Mon. Not. R. Astron. Soc.* **341**, 54 (2003).
[17] Wiedner, C., Kroupa, P., & Larsen, S. *Mon. Not. R. Astron. Soc.* **350**, 1503 (2004).
[18] See, e.g., Eldridge, J. J., & Stanway, E. R. *Mon. Not. R. Astron. Soc.* **400**, 1019 (2009); Raiter, A., Schaerer, D., & Fosbury, R.A.E. *Astron. & Astrophys.* **523**, A64 (2010).
[19] Robertson, B., et al. *Nature* **468**, 49 (2010).
[20] Obreschkow, D., et al. *Astrophys. J.* **702**, 1321 (2009).
[21] See, e.g., Shapley, A. E., et al. *Astrophys. J.* **588**, 65 (2003).
[22] Wang, R., et al. *Astrophys. J.* **714**, 699 (2010).
[23] Krumholz, M. R., Leroy, A. K., & McKee, C. F. *Astrophys. J.* **731**, 25 (2011); Genzel, R., et al. *Nature*, in press, arXiv.org/1106.2098 (2011).
[24] Bradley, L., et al. *Astrophys. J.* **678**, 647 (2008).

Chapter 9

[1] Strömgren, B. *Astrophys. J.* **89**, 526 (1939).
[2] Shapiro, P. R., & Giroux, M. L. *Astrophys. J.* **321** L107 (1987).
[3] Wyithe, J.S.B., & Loeb,A. *Nature* **427**, 815 (2004); *Astrophys. J.* **610**, 117 (2004).
[4] Mellema, G., et al. *Mon. Not. R. Astron. Soc.* **372**, 679 (2006).
[5] See Figure 12 in Trac, H., & Cen, R. *Astrophys. J.* **671**, 1 (2007), Pawlik, A. H., et al. *Mon. Not. R. Astron. Soc.* **394**, 1812 (2009).
[6] See, e.g., Nestor, D. B., et al. *Astrophys. J.* **736**, 18 (2011), and references therein.
[7] Furlanetto, S. R., Zaldarriaga, M., & Hernquist, L. *Astrophys. J.* **613**, 1 (2004).
[8] Wyithe, J.S.B., & Loeb, A. *Nature* **432**, 7014 (2004); Barkana,R., & Loeb, A. *Mon. Not. R. Astron. Soc.* **372**, 43 (2006).
[9] Barkana, R., & Loeb, A. *Astrophys. J.* **609**, 474 (2004).
[10] Wyithe, J.S.B. & Loeb, A. *Nature* **432**, 194. (2004).
[11] Sheth, R. K. *Mon. Not. R. Astron. Soc.* **300**, 1057 (1998).
[12] Maggiore, M., & Riotto, A. *Astrophys. J.* **711**, 907 (2010).
[13] Choudhury, T. R., Haehnelt, M., & Regan, J. *Mon. Not. R. Astron. Soc.* **394**, 960 (2009); Crociani, D., et al. *Mon. Not. R. Astron. Soc.*, **411**, 289 (2011).
[14] Osterbrock, D. E., & Ferland, G. J. *Astrophysics of Gaseous Nebulaeand Active Galactic Nuclei*, 2nd ed. Mill Valley, CA: University Science Books (2006).
[15] See, e.g., Gnedin, N. Y. *Astrophys. J.* **535**, 530 (2000); Shapiro, P. R., Iliev, I. T., & Raga,A. C. *Mon. Not. R. Astron. Soc.* **348**, 753 (2004).
[16] Iliev, I., et al. *Mon. Not. R. Astron. Soc.* **371**, 1057 (2006); **400**, 1283 (2009).
[17] McQuinn, M., et al. *Mon. Not. R. Astron. Soc.* **377**, 1043 (2007).

[18] See, e.g., Zahn, O., et al. *Astrophys. J.* **654**, 12 (2007); Mesinger, A., & Furlanetto, S. R. *Astrophys. J.* **669**, 663 (2007); Mesinger, A., Furlanetto, S. R., & Cen, R. *Mon. Not. R. Astron. Soc.* **411**, 955 (2011).
[19] Zahn, O., et al. *Mon. Not. R. Astron. Soc.*, **414**, 727 (2011).
[20] Barkana, R. *Mon. Not. R. Astron. Soc.* **376**, 1784 (2007).
[21] McQuinn, M., et al. *Astrophys. J.* **630**, 643 (2005).
[22] Fan, X., et al. *Astron. J.* **122**, 2833(2001).
[23] Hickox, R. C., & Markevitch, M. *Astrophys. J.* **671**, 1523 (2007).
[24] Dijkstra, M., Haiman, Z., & Loeb, A. *Astrophys. J.* **613**, 646 (2004).
[25] Mirabel, I. F., et al. *Astron. & Astrophys.* **528**, 149 (2011); Dijkstra, M., Gilfanov, M., Loeb, A., & Sunyaev, R. *Mon. Not. R. Astron. Soc.* **421**, 213 (2012).
[26] Furlanetto, S. R., & Johnson Stoever, S. *Mon. Not. R. Astron. Soc.* **404**, 1869 (2010).
[27] Furlanetto, S. R. & Johnson Stoever, S. *Mon. Not. R. Astron. Soc.* **404**, 1869 (2010).
[28] Wyithe, J.S.B., & Loeb, A. *Astrophys. J.* **586**, 693 (2003).
[29] Furlanetto, S. R., & Oh, S. P. *Astrophys. J.* **681**, 1 (2008).
[30] Dixon, K. L., & Furlanetto, S. R. *Astrophys. J.* **706**, 970 (2009).
[31] Shull, J. M., et al. *Astrophys. J.* **722**, 1312 (2010); Worseck, G., et al. *Astrophys. J. Lett.* **733**, L24 (2011).
[32] Schaye, J., et al. *Mon. Not. R. Astron. Soc.* **318**, 817 (2000); Lidz, A., et al. *Astrophys. J.* **718**, 199 (2010); Becker, G. D., et al. *Mon. Not. R. Astron. Soc.* **410**, 1096 (2011).
[33] See, e.g., Songaila, A. *Astron. J.* **130**, 1996 (2005); Aguirre, A., et al. *Astrophys. J.* **602**, 38 (2004).
[34] See, e.g., Furlanetto, S. R., Oh, S. P., & Pierpaoli, E. *Phys. Rev. D* **74**, 103502 (2006).
[35] Shapiro, P. R., Iliev, I. T., & Raga, A. C. *Mon. Not. R. Astron. Soc.* **348**, 753 (2004).
[36] Iliev, I. T., Shapiro, P. R., & Raga, A. C. *Mon. Not. R. Astron. Soc.* **361**, 405 (2005).
[37] Iliev, I. T., Scannapieco, E. S., & Shapiro, P. R. *Astrophys. J.* **624**, 491 (2005).
[38] Oh, S. P., & Haiman, Z. *Mon. Not. R. Astron. Soc.* **346**, 456 (2003).
[39] Dijkstra, M., et al. *Astrophys. J.* **601**, 666 (2004).

Chapter 10

[1] Turner, E. private communication (2009).
[2] http://www.eso.org/sci/facilities/eelt/.
[3] http://www.gmto.org/.
[4] http://www.tmt.org/.
[5] http://almaobservatory.org/.
[6] Stark, D., et al. *Astrophys. J.* **663**, 10 (2007); Bouwens, R. J., et al. *Astrophys. J* **690**, 1764 (2009).

[7] Richard, J., et al. *Astrophys. J.* **685**, 705 (2008).
[8] Bradley, L. D. et al. *Astrophys. J.* **747**, 3 (2012); Schenker, M. A. et al., *Astrophys. J.* **744**, 179 (2012).
[9] Cucchiara, A., et al. *Astrophys. J.* **736**, 7 (2011).
[10] Ciardi, B., & Loeb, A. *Astrophys. J.* **540**, 687 (2000).
[11] See Ellis, R. S. in *First Light in the Universe*, SAAS-Fee Advanced Course **36**, Springer, New York (2008), and references therein, for more information.
[12] Tegmark, M. *Phys. Rev. Lett.* **79**, 3806 (1997).

Chapter 11

[1] Partridge, R. B., & Peebles, P.J.E. *Astrophys. J.* **147**, 868 (1967).
[2] Schaerer, D. *Astr. & Astrophys.* **397**, 527 (2003).
[3] Hansen, M., & Oh, S. P. *Mon. Not. R. Astron. Soc.* **367**, 979 (2006).
[4] Neufeld, D. A. *Astrophys. J.* **370**, L85 (1991).
[5] Miralda-Escudé, J. *Astrophys. J.* **501**, 15 (1998).
[6] Chen, H.-W., Prochaska, J. X., & Gnedin, N. Y. *Astrophys. J. Lett.* **667**, 125 (2007).
[7] Mesinger, A. *Mon. Not. R. Astron. Soc.* **407**, 1328 (2010).
[8] See also Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **386**, 492 (2008).
[9] Steidel, C. C., et al. *Astrophys. J.* **532**, 170 (2000).
[10] Matsuda, Y., et al. *Astron. J.*, **128**, 569 (2004); *Astrophys. J.* **640**, L123 (2006); Saito, T., et al. *Astrophys. J.* **537**, L5 (2000); Yang, Y., et al. *Astrophys. J.* **693**, 1579 (2009).
[11] 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); Geach, J. E., et al. *Astrophys. J.* **655**, L9(2007); Basu-Zych, A., & Scharf, C. *Astrophys. J.* **615**, L85 (2004).
[12] Nilsson, K. K., et al. *A & A.* **452**, L23 (2006); Smith, D.J.B., et al. *Mon. Not. R. Astron. Soc.*, **378**, L49 (2007).
[13] Haiman, Z., Spaans, M., & Quataert, E. *Astrophys. J.* **537**, L5 (2000); Fardal, M. A., et al. *Astrophys. J.* **562**, 605 (2001).
[14] Haiman, Z., & Rees, M. J. *Astrophys. J.* **556**, 87 (2001).
[15] Scharf, C. et al. *Astrophys. J.*, **596**, 105 (2003).
[16] Mori., M., et al. *Astrophys.J.* **613**, L97 (2004).
[17] Rees, M. J. *Mon. Not. R. Astron. Soc.*, **239**, 1P (1989).
[18] Dijkstra, M., & Loeb, A. *Mon. Not. R. Astron. Soc.* **400**, 1109 (2009); Goerdt, T., et al. *Mon. Not. R. Astron. Soc.* **407**, 613 (2010); Faucher-Giguére, C.-A., et al. *Astrophys. J.* **725**, 633 (2010).
[19] Keres, D., et al. *Mon. Not. R. Astron. Soc.*, **363**, 2 (2005); Keres, D., et al. *Mon. Not. R. Astron. Soc.* **395**, 160 (2009).
[20] Gould, A., & Weinberg, D. H. *Astrophys. J.* **468**, 462 (1996).
[21] Cantalupo, S., Porciano, C., & Lilly, S. J. *Astrophys. J.* **672**, 48 (2008).
[22] Zheng, Z., et al. *Astrophys. J.* **726**, 38 (2011).

Chapter 12

[1] Rybicki, G. B., & Lightman, A. P. *Radiative Processesin Astro physics*, New York: Wiley (1986).
[2] Field, G. B. *Astrophys. J.* **129**, 525 (1959).
[3] Zygelman, B., *Astrophys. J.* **622**, 1356 (2005); Furlanetto, S. R., & Furlanetto, M. R. *Mon. Not. R. Astron. Soc.* **374**, 547 (2007); **379**, 130 (2007).
[4] Wouthuysen, S. A. *Astron. J.* **57**, 31 (1952); Field, G. B.*Proc. Inst. Radio Eng.* **46**, 240 (1958).
[5] Pritchard, J. R., & Furlanetto, S. R. *Mon. Not. R. Astron. Soc.* **367**, 1057 (2006).; Hirata, C. M. *Mon. Not. R. Astron. Soc.* **367**, 259 (2006).
[6] See, e.g., Bethe, H. A., & Salpeter, E. E. *Quantum Mechanics of One-and Two-Electron Atoms*. New York: Academic Press (1957).
[7] Field, G. B. *Astrophys. J.* **129**, 551 (1959).
[8] Ibid.
[9] Hirata, C. M. *Mon. Not. R. Astron. Soc.* **367**, 259 (2006).
[10] Ibid.
[11] Dalal, N., Pen,U. L., & Seljak, U. *J. Cos. Astropart. Phys.* **11**, 007 (2010).
[12] Barkana, R., & Loeb, A. *Astrophys. J.* **626**, 1 (2005).
[13] Furlanetto, S. R., & Johnson Stoever, S. *Mon. Not. R. Astron. Soc.* **404**, 1869 (2010).
[14] Furlanetto, S. R., & Pritchard, J. R. *Mon. Not. R. Astron. Soc.* **372**, 1093 (2006).
[15] Mirabel, F., et al. *Astron. & Astrophys.* **528**, 149 (2011).
[16] See http://www.haystack.mit.edu/ast/arrays/Edges/.
[17] Bowman, J. D. & Rogers, A.E.E. *Nature* **468**, 796 (2010).
[18] Kaiser N., *Mon. Not. R. Astron. Soc.* **227**, 1 (1987).
[19] Barkana, R., & Loeb, A. *Astrophys. J.* **624**, 65 (2005).
[20] Kanekar, N., & Briggs, F. H. *New Astronomy Reviews* **48**, 1259 (2004).
[21] Mao, Y., et al. *Phys. Rev. D* **78**, 023529 (2008).
[22] Furlanetto, S. R., Oh, S. P., & Briggs, F. H. *Phys. Rep.* **433**, 181 (2006).
[23] Paciga, G., et al. *Mon. Not. R. Astron. Soc.* **413**, 1174 (2011).
[24] See, e.g., Dodelson, S. *Modern Cosmology*, Amsterdam: Academic Press (2003) for a pedagogical summary.

Chapter 13

[1] See a similar diagram and complementary discussion in Kamionkowski, M., & Loeb, A. *Phys. Rev.* **D56**, 4511 (1997).
[2] Zaldarriaga, M. *Phys. Rev.* **D44**, 1822 (1997).
[3] Sunyaev, R. A., & Zel'dovich, I. B. *Mon. Not. R. Astron. Soc.* **190**, 413 (1980).
[4] Dunkley, J., et al. *Astrophys. J. Supp.* **180**, 306 (2009).
[5] See http://www.rssd.esa.int/index.php?project=planck.
[6] Mortonson, M. J., & Hu, W. *Phys. Rev.* **D77**, 043506 (2008).
[7] Holder, G. P., Iliev, I. T., & Mellema, G. *Astrophys. J. Lett.* **663**, L1 (2007).
[8] See, e.g., Dodelson, S. *Modern Cosmology*. New York: Academic Press (2003).
[9] Jaffe, A., & Kamionkowski, M. *Phys. Rev.* **D58**, 043001 (1998).

[10] Ostriker, J. P., & Vishniac, E. T. *Astrophys. J.* **306**, L51(1986); Vishniac, E. T. *Astrophys. J.* **322**, 597 (1987).
[11] Gruzinov, A. & Hu, W. *Astrophys. J.* **508**, 435 (1998); Knox, L., Scoccimarrao, R., & Dodelson, S. *Phys. Rev. Lett.* **81**, 2004 (1998).
[12] Dunkley, J., et al. *Astrophys. J.* **739**, 52 (2011); Reichardt, C. L., et al. *Astrophys. J.*, submitted, arXiv.org/1111.0932 (2011).
[13] Fixsen, D. J., et al. *Astrophys. J.* **473**, 576 (1996).
[14] Tegmark, M., Silk, J., & Evrard, A., *Astrophys. J.* **417**, 54 (1993).
[15] Fernandez, E. R., & Komatsu, E. *Astrophys. J.* **646**, 703 (2006).
[16] Rybicki, G. B., & Lightman, A. P. *Radiative Processes in Astrophysics*. NewYork: Wiley (1979), pp. 160–161.
[17] Brown, R. L., & Mathews, W. G. *Astrophys. J.* **160**, 939 (1970).
[18] Kashlinksky, A., et al. *Astrophys. J. Lett.* **654**, L1 (2007); Thompson, R. I., et al. *Astrophys. J.* **657**, 669 (2007); Cooray, A., et al. *Astrophys. J. Lett.* **659**, L91 (2007).
[19] Kashlinsky, A. et al. *Astrophys. J.* **654**, L5 (2007); Cooray, A., et al. *Astrophys. J.* **659**, L91 (2007).
[20] Thompson, R. I., et al. *Astrophys. J.* **657**, 669 (2007).
[21] Lidz, A., et al. *Astrophys. J.* **659**, 865 (2007).
[22] Alvarez, M. A., et al. *Astrophys. J.* **647**, 840 (2006).
[23] Adshead, P. J. & Furlanetto, S. R. *Mon. Not. R. Astron. Soc.* **384**, 291 (2008).
[24] Jelic, V., et al. *Mon. Not. R. Astron. Soc.* **402**, 2279 (2010).
[25] L. Gao, et al. *Mon. Not. R. Astron. Soc.* **403**, 1283 (2010).
[26] Scannapieco, E., et al. *Astrophys. J.* **653**, 285 (2006).
[27] Tumlinson, J. *Astrophys. J.* **708**, 1398 (2010).
[28] Sellwood, J. A., & Binney, J. J. *Mon. Not. R. Astron. Soc.* **336**, 785 (2002).
[29] See, e.g., Heger, A., & Woosley, S. E., *Astrophys. J.* **567**, 532 (2002).
[30] Gal-Yam, A., et al. *Nature* **462**, 624 (2009).
[31] Cooke, R., et al. *Mon. Not. R. Astron. Soc.* **412**, 1047 (2011).
[32] See, e.g., Karlsson, T., Bromm, V., & Bland-Hawthorn, J., arXiv.org/1101.4024 (2011), and references therein.

Further Reading

The following books, review articles, and other resources expand on elements of the text in each of the chapters of this book. We refer the reader to the Notes for references to specific papers in the astrophysical literature.

Chapter 1: Introduction and Cosmological Background, and Chapter 2: Linear Growth of Cosmological Perturbations

Dodelson, S. *Modern Cosmology*, Academic Press (2003).
Kolb, E. W., & Turner, M. S. *The Early Universe*, Addison Wesley (1990).
Liddle, D., & Lyth, A. *The Primordial Density Perturbation: Cosmology, Inflation, and the Origin of Structure*, Cambridge University Press (2009).
Mukhanov, V. *Physical Foundations of Cosmology*, Cambridge University Press (2005).
Padmanabhan, T. *Structure Formation in the Universe*, Cambridge University Press (1993).
Peacock, J. A. *Cosmological Physics*, Cambridge University Press (1998).
Peebles, P.J.E. *Principles of Physical Cosmology*, Princeton University Press (1993).
Schneider, P. *Extragalactic Astronomy and Cosmology*, Springer-Verlag (2006).
Other Resources: Two useful computer codes are *CAMB* (http://camb.info/), used for generating CMB anisotropies as well as accurate linear power spectra, and *RECFAST* (http://www.astro.ubc.ca/people/scott/recfast.html), for calculating the recombination history of the Universe.

Chapter 3: Nonlinear Structure and Halo Formation

Cooray, A., & Sheth, R. Halo Models of Large-Scale Structure. *Physics Reports* **382**, 1 (2002).
Springel, V. Smoothed Particle Hydrodynamics in Astrophysics. *Annual Reviews of Astronomy & Astrophysics* **48**, 391 (2010).
Zentner, A. R. The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth. *International Journal of Modern Physics D* **16**, 763 (2007).
Other Resources: Two publicly available numerical simulation codes are the *N*-body/SPH code *GADGET* (http://www.mpa-garching.mpg.de/galform/gadget/) and the *N*-body/adaptive mesh refinement code *Enzo* (http://lca.ucsd.edu/portal/software/enzo).

Chapter 4: The Intergalactic Medium

Fan, X., Carillii, C.L., & Keating, B. Observational Constraints on Cosmic Reionization. *Annual Reviews of Astronomy & Astrophysics*. **44**, 415 (2006).

Meiksin, A. The Physics of the Intergalactic Medium. *Reviews of Modern Physics* **81**, 1405 (2009).

Osterbrock, D. E., & Ferland, G. J. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*, 2nd ed. University Science Books (2006).

Rybicki, G. B., & Lightman, A. P. *Radiative Processes in Astrophysics*. Wiley-Interscience (1979).

Chapter 5: The First Stars

Bromm, V., & Larson, R. B. The First Stars. *Annual Reviews of Astronomy & Astrophysics*. **42**, 79 (2004).

Kippenhahn, R., & Weigert, A. *Stellar Structure and Evolution*. Springer-Verlag (1994)

Loeb, A. *How Did the First Stars and Galaxies Form?* Princeton University Press (2010)

Piran, T. The Physics of Gamma-Ray Bursts. *Reviews of Modern Physics* **76**, 1143 (2004).

Shull, J. M., & Beckwith, S. Interstellar Molecular Hydrogen. *Annual Reviews of Astronomy & Astrophysics* **20**, 163 (1982).

Chapter 6: Stellar Feedback and Galaxy Formation

Bromm, V., & Yoshida, N. The First Galaxies. *Annual Reviews of Astronomy & Astrophysics* **49**, 373 (2011).

Ciardi, B., & Ferrara, A. The First Cosmic Structures and Their Effects. *Space Science Reviews* **116**, 625 (2005).

Mo, H., van den Bosch, F., & White, S. *Galaxy Formation and Evolution*, Cambridge University Press (2010).

Chapter 7: Supermassive Black Holes

Colpi, M., & Dotti, M. *Massive Binary Black Holes in the Cosmic Landscape*. http://arxiv.org/abs/0906.4339v1 (2009).

Peterson, B. M. *An Introduction to Active Galactic Nuclei*. Cambridge University Press (1997).

Shapiro, S. L., & Teukolsky, S. A. *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects*. Wiley-Interscience (1983).

Volonteri, M. Formation of Supermassive Black Holes. *Astronomy and Astrophysics Review* **18**, 279 (2010).

Chapter 8: Physics of Galaxy Evolution

Benson, A. J. Galaxy Formation Theory. *Physics Reports* **495**, 33 (2010).

Binney, J., & Merrifield, M. Galactic Astronomy, Princeton University Press (1998).

Binney, J., & Tremaine, S. *Galactic Dynamics*, 2nd ed. Princeton University Press (2008).
Draine, B. T. *Physics of the Interstellar and Intergalactic Medium*. Princeton University Press (2011).
Mo, H., van den Bosch, F., & White, S. *Galaxy Formation and Evolution*. Cambridge University Press (2010).
Schneider, P., Ehlers, J., & Falco, E. E. *Gravitational Lenses*. Springer (1999).
Other Resources: Publicly available stellar population synthesis codes include *BPASS* (http://www.bpass.org.uk/) and *GALAXEV* (http://www.cida.ve/∼bruzual/bc2003).

Chapter 9: The Reionization of Intergalactic Hydrogen

Barkana, R., & Loeb, A. In the Beginning: The First Sources of Light and the Reionization of the Universe. *Physics Reports* **349**, 125 (2001).
Barkana, R., & Loeb, A. The Physics and Early History of the Intergalactic Medium. *Reports on Progress in Physics* **70**, 627 (2007).
Other Resources: Publicly available seminumerical reionization codes include *DexM* and *21cmFAST* (both available from http://homepage.sns.it/mesinger/DexM___21cmFAST.html)

Chapter 10: Surveys of High-Redshift Galaxies

Binney, J., & Merrifield, M. *Galactic Astronomy*. Princeton University Press (1998).
Dodelson, S. *Modern Cosmology*. Academic Press (2003).
Ellis, R. S. Observations of the High-Redshift Universe. First Light in the Universe, Saas-Fee Advanced Courses **36**, 259 (2008).
Mo, H., van den Bosch, F., & White, S. *Galaxy Formation and Evolution*. Cambridge University Press (2010).

Chapter 11: The Lyman-α Line as a Probe of the Early Universe

Draine, B. T. *Physics of the Interstellar and Intergalactic Medium*, Princeton University Press (2011).

Chapter 12: The 21-cm Line

Furlanetto, S. R., Oh, S. P., & Briggs, F. H. Cosmology at Low Frequencies: The 21-cm Transition and the High-Redshift Universe. *Physics Reports* **433**, 181 (2006).
Morales, M. F., & Wyithe, J.S.B. Reionization and Cosmology with 21-cm Fluctuations. *Annual Reviews of Astronomy & Astrophysics* **48**, 127 (2010).
Pritchard, J., & Loeb, A. 21-cm Cosmology. *Reports on Progress in Physics*, http://arxiv.org/abs/1109.6012 (2012).
Thompson, R. A., Moran, J. M., & Swenson, G. W. Jr. *Interferometry and Synthesis in Radio Astronomy*, 2nd ed. Wiley-Interscience (2001).

Other Resources: A publicly available seminumerical code for computing the 21-cm background is *21cmFAST* (http://homepage.sns.it/mesinger/DexM___21cmFAST.html).

Chapter 13: Other Probes of the First Galaxies

Aghanim, N., Majumdar, S., & Silk, J. Secondary Anisotropies of the CMB. *Reports on Progress in Physics* **71**, 066902 (2008).

Birkinshaw, M. The Sunyaev-Zel'dovich effect. *Physics Reports* **310**, 97 (1999).

Dodelson, S. *Modern Cosmology*. Academic Press (2003).

Frebel, A., "Exploring the Universe with Metal-Poor Stars," to appear in *The First Galaxies: Theoretical Predictions and Observational Clues*, ed. V. Bromm, B. Mobasher, & T. Wiklind. Springer, arXiv.org/1108.4692 (2012).

Hu, W., & White, M. A CMB Polarization Primer. *New Astronomy* **2**, 323 (1997).

Kashlinsky, A. Cosmic Infrared Background and Early Galaxy Evolution. *Physics Reports* **409**, 361 (2005).

Karlsson, T., Bromm, V., & Bland-Hawthorn, J. Pre-galactic Metal Enrichment: The Chemical Signatures of the First Stars. *Reviews of Modern Physics*, http://arxiv.org/abs/1101.4024 (2012).

Index begins new recto, page 513

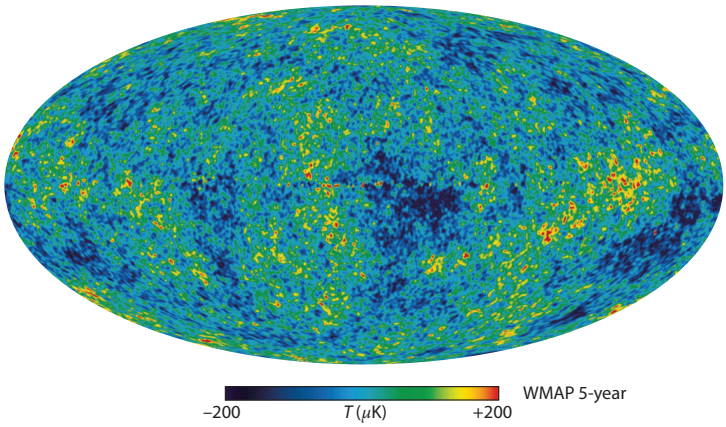


Plate 1 Image of the Universe when it first became transparent, 400,000 years after the Big Bang, taken over 5 years by the *Wilkinson Microwave Anisotropy Probe* (WMAP) satellite (see Figure 1.1 and associated discussion for more information). Slight density inhomogeneities at the level of one part 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 microkelvins, and the unperturbed temperature is 2.725 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. NASA and the WMAP Science Team. Courtesy of /

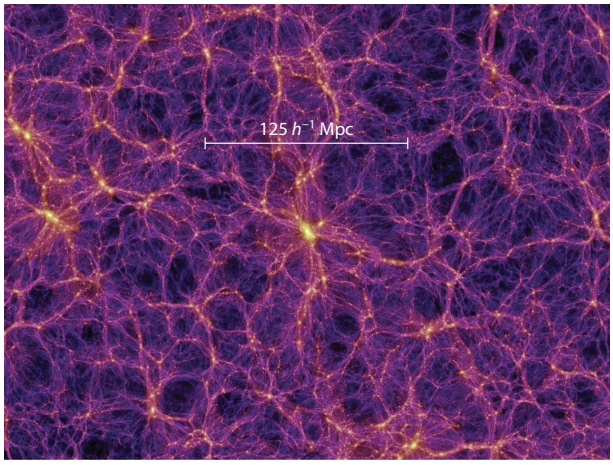


Plate 2 Slice through the *Millennium Simulation*, a massive computer simulation of cosmological structure formation (see Figure 4.1 and associated discussion for more information). The color scale shows the dark matter density; note how matter is organized into dense filaments (in many cases, these are actually slices through sheets of matter) separating nearly empty voids. Massive galaxies and galaxy clusters form at the intersections of these filaments. V. Springel (2005). Courtesy of /

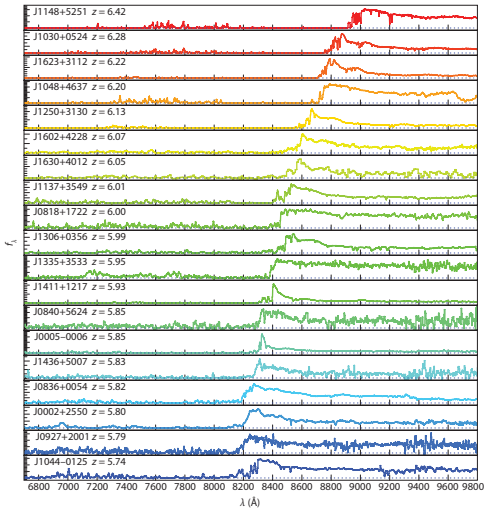


Plate 3 Observed spectra of 19 quasars with redshifts $5.74 < z < 6.42$ from the Sloan Digital Sky Survey (see Figure 4.6 and associated discussion for more information). For some of the highest-redshift quasars, the spectra show no transmitted flux shortward of the Lyman-α wavelength at the quasar redshift, 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 broadband photometry is adequate for inferring the redshift of sources during the epoch of reionization. Fan, X., et al., *Astron. J.* ~~128, 515 (2004).~~ Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

132, 117 (2006). (132 is bold, remainder plain) / AA

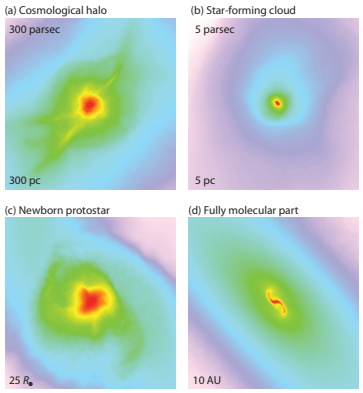


Plate 4 Projected gas distribution around a primordial protostar from a numerical simulation (see Figure 5.6 and associated discussion for more information). Shown is the gas density of a single object on different spatial scales: *(a)* the large-scale gas distribution around the cosmological minihalo; *(b)* the self-gravitating, star-forming cloud; *(c)* the central part of the fully molecular core; and *(d)* the final protostar. ~~Yoshida, N., Omukai, K., & Hernquist, L., *Science* **321**, 669 (2008). Copyright 2008 by the American Association for the Advancement of Science.~~

Reprinted from *Science*, 321/5889, Yoshida, N., K. Omukai, & L. Hernquist, "Protostar Formation in the Early Universe," <~?~AU: add page number>, Copyright 2008, with the permission of The American Association for the Advancement of Science.

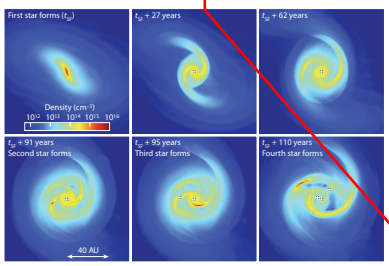


Plate 5 Density evolution in a 120 AU region around the first protostar in a numerical simulation of Population III star formation, showing the buildup of the protostellar disk and its eventual fragmentation at the times labeled in the diagram (see Figure 5.7 and associated discussion for more information). ~~Clark, P. C., et al., *Science* **331**, 1040 (2011). Copyright 2011 by the American Association for the Advancement of Science.~~

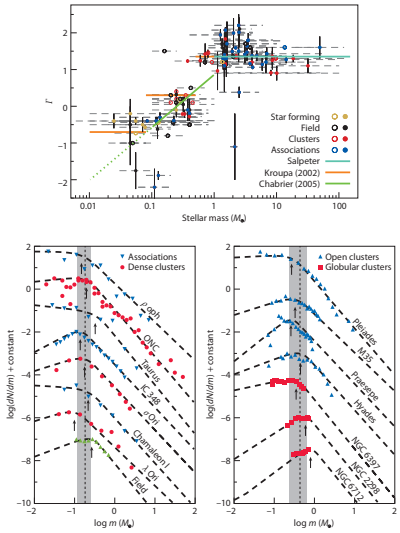


Plate 6 *Upper panel:* The derived power-law index, Γ, of the IMF in nearby star-forming regions, clusters, and associations of stars within the Milky Way galaxy, as a function of sampled stellar mass (points are placed in the center of the log $m_\star$ range used to derive each index; the dashed lines indicate the full range of masses sampled). The colored solid lines represent three analytic 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, the dotted line indicates the mean characteristic mass among the clusters in each panel, and the shaded region shows the standard deviation of the characteristic masses in that panel. The observations are consistent with a single underlying IMF. (see Figure 5.9 and associated discussion for more information) Bastian, N., Covey, K. R., & Meyer, M. R., *Ann. Rev. Astr. & Astrophys.* **48** (2010). Copyright 2010 by Annual Reviews.

Capitalize S / PE

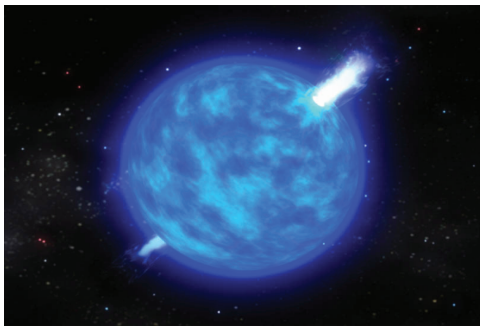


Plate 7 Illustration of a long-duration gamma-ray burst in the popular collapsar model (see Figure 5.18 and associated discussion for more information). The collapse of the core of a massive star (which lost its hydrogen envelope) to a black hole generates two opposite jets moving out at a speed close to that of light. The jets drill a hole in the star and shine brightly toward 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. Courtesy of / NASA E/PO.

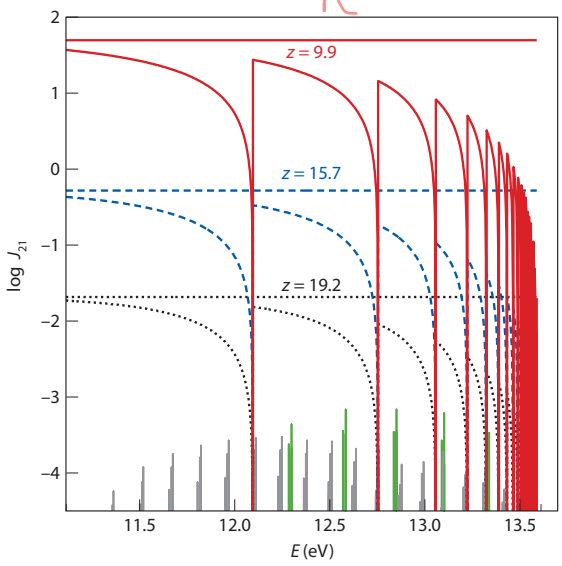


Plate 8 The "sawtooth" modulation of a uniform, spectrally flat radiation background in the Lyman-Werner frequency band when the IGM is still predominantly neutral (see Figure 6.2 and associated discussion for more information). The three curves are for $z = 19.2$, 15.7, and 9.2, from top to bottom; the horizontal lines show the unattenuated spectrum, while the curves with features show the effect of Lyman-series absorption. The vertical lines at the bottom of the figure show some of the Lyman-Werner transitions, with the height equal to 1% of the oscillator strength. Ahn, K., et al., *Astrophys. J.* **695**, 1430 (2009). Reproduced by / AAS. / ~~with~~ permission of the ~~American Astronomical Society.~~

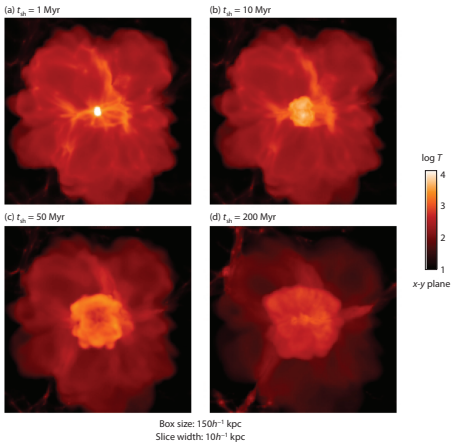


Plate 9 Temperature maps from a numerical simulation of a supernova explosion (see Figure 6.7 and associated discussion for more information). The supernova of a 200 $M_\odot$ star is set off at $z \approx 20$ in a halo with $M = 5 \times 10^5$ $M_\odot$ and $r_{\rm vir} \approx 100$ pc. The snapshots are 1, 10, 50, and 200 million years after the explosion. In the first panel on the top left, the supernova is the central hot region; the star's H II region fills most of the box (fading with time as the gas recombines and cools). The supernova remnant expands over the four panels, gradually becoming more anisotropic as it encompasses the filamentary structure surrounding the halo. Greif, T., et al., *Astrophys. J.* **670**, 1 (2007). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

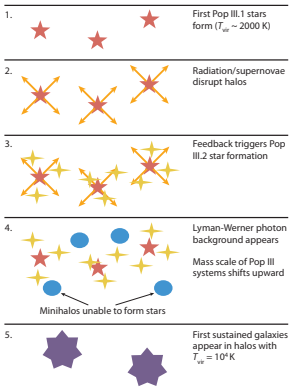
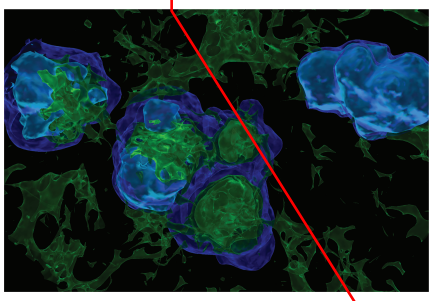


Plate 10 Stages in a plausible scenario for the birth of the first stars and galaxies (see Figure 6.12 and associated discussion for more information). *(1)* The first Population III.1 stars form in small halos via H_2 cooling. *(2)* These stars empty their hosts of gas via photoevaporation and supernova blast waves. *(3)* This feedback triggers Population III.2 star formation in nearby minihalos. *(4)* The Lyman-Werner background from these stars suppresses star formation in small minihalos, gradually increasing the characteristic mass scale of star-forming objects. *(5)* The first self-sustaining galaxies eventually form in halos above the atomic cooling threshold, $T_{\rm vir} \sim 10^4$ K.

Reprinted from *Nature*, 459/7243, Bromm, V., N. Yoshida, L. Hernquist, & C.F. McKee, "The formation of the first stars and galaxies," Figure 4, Copyright 2007, with permission from Nature Publishing Group.



blue / AA

green / AA

Plate 11 Results from a numerical simulation of the formation of a metal-free stars and their feedback on the surrounding environment (see Figure 6.13 and associated discussion for more information). Radiative feedback around the first star involves ionized bubbles (light gray) and regions of high molecule abundance (medium gray). The large residual free-electron fraction inside the relic ionized regions, left behind after the central star died, rapidly catalyzes the re-formation of molecules and a new generation of lower-mass stars. Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F., *Nature* **459**, 49 (2009). Copyright 2009 by Nature Publishing Group.

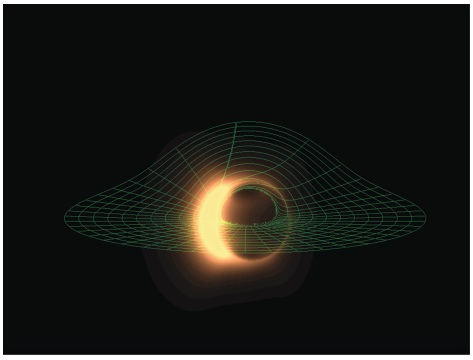


Plate 12 Simulated image of an accretion flow around a black hole spinning at half its maximum rate, from a viewing angle of 10° relative to the rotation axis (see Figure 7.2 and associated discussion for more information). The coordinate grid in the equatorial plane of the spiraling flow shows how strong lensing around the black hole bends the back of the apparent disk up. The left side of the image is brighter owing to its rotational motion toward 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, interferometers operating at millimeter wavelengths across the earth are necessary. Broderick, A., & Loeb, A. *Journal of Physics Conf. Ser.* **54**, 448 (2006); *Astrophys. J.* **697**, 1164 (2009). Reproduced with permission of the American Astronomical Society.

by / AAS. /

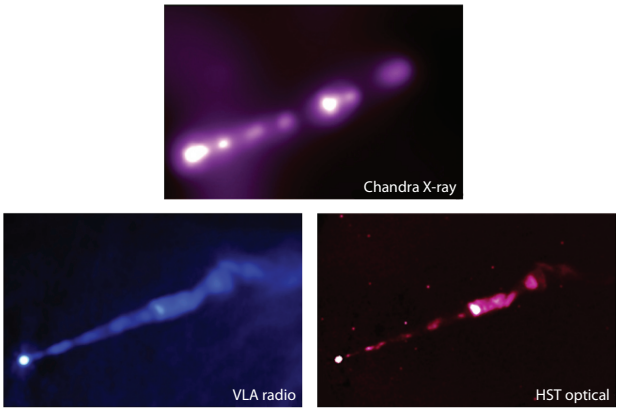


Plate 13 Multiwavelength images of the highly collimated jet emanating from the supermassive black hole at the center of the giant elliptical galaxy M87 (see Figure 7.3 and associated discussion for more information). The X-ray image (*top*) was obtained with the Chandra X-ray observatory, the radio 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). Courtesy of / CXO/NASA.

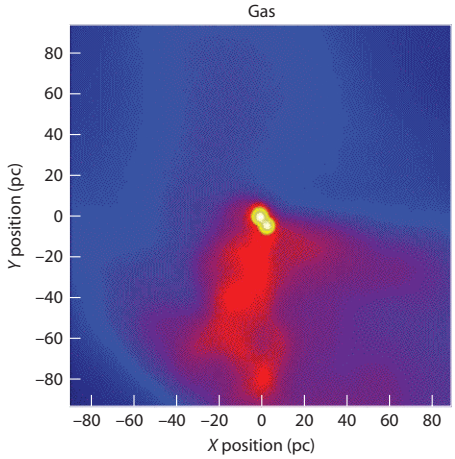


Plate 14 Numerical simulation of the collapse of an early dwarf galaxy with a virial temperature just above the cooling threshold of atomic hydrogen and no H_2 (see Figure 7.6 and associated discussion for more information). The image shows a snapshot of the gas density distribution 500 million years 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. Subfragmentation 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 for supermassive black holes. The simulated box size is 200 pc on a side. Bromm, V., & Loeb, A., *Astrophys. J.* **596**, 34 (2003). Reproduced by / ~~with~~ permission of the ~~American Astronomical Society.~~ AAS. /

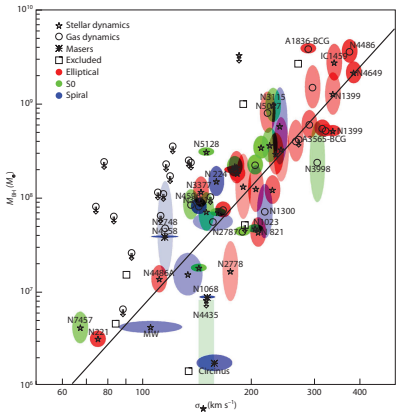


Plate 15 Dynamical measurements of the correlation between supermassive black hole mass, $M_{\rm BH}$, and velocity dispersion of stars in the spheroid of its host galaxy, $\sigma_\star$ (see Figure 7.7 and associated discussion for more information). The symbol indicates the method of black hole mass measurement: dynamics of stars (*pentagrams*), dynamics of gas (*circles*), dynamics of maser sites (*asterisks*). Arrows indicate 3σ upper limits to black hole mass. The shade of the error ellipse indicates the Hubble type of the host galaxy: elliptical, S0, or spiral. The line is the best-fit relation to the full sample: $M_{\rm BH} = 10^{8.12}\, M_\odot (\sigma_\star/200\,{\rm km\,s^{-1}})^{4.24}$. The mass uncertainty for NGC 4258 has been plotted much larger than its actual value to show on this plot. Gültekin, K., et al., *Astrophys. J.* **698**, 198 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /

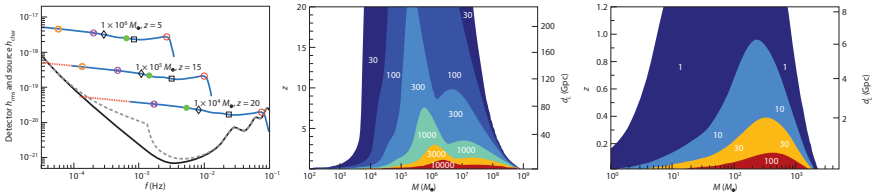


Plate 16 Sensitivity of the future gravitational wave observatories, LISA and Advanced-LIGO, to equal-mass ($M_1 = M_2 = M/2$) binaries (see Figure 7.8 and associated discussion for more information). *Left panel:* Root-mean-square noise amplitude of LISA $h_{\rm rms}$ from the detector only (dashed) and from the detector combined with the anticipated foreground confusion (dash-dotted), along with the characteristic amplitudes $h_{\rm char}$ of three binary masses M (solid). The locations on each $h_{\rm char}$ curve correspond to the peak amplitude (circle), 1 hour before the peak (filled circle), 1 day before the peak (circle with inscribed cross), and 1 month before the peak (circle with inscribed square) in the observer frame, as well as times of $25 r_{\rm Sch}/c$ (square) and $500 r_{\rm Sch}/c$ (diamond) before the peak in the source frame. *Middle panel:* Contour plot of the signal-to-noise ratio (SNR) with binary mass and redshift dependence for LISA. *Right panel:* SNR contour plot with mass and redshift dependence for Advanced-LIGO. Baker, J., et al., *Phys. Rev.* **D75**, 4024 (2007). Copyright 2007 by the American Physical Society.

Reprinted Figures 16, 17, and 19 from / AA

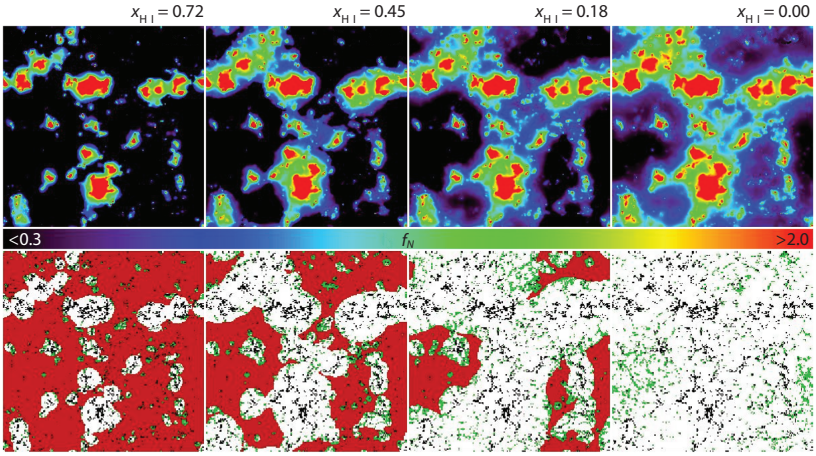
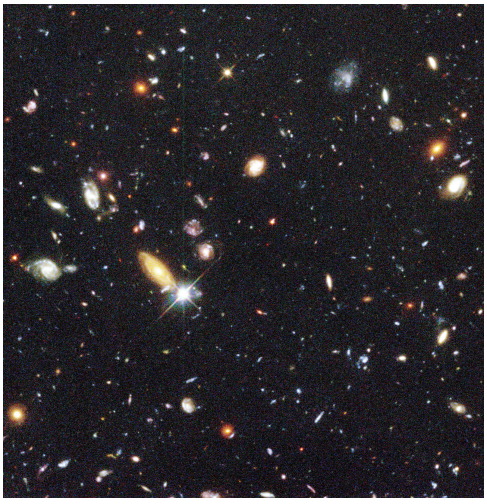


Plate 17 IGM absorbers in a "seminumerical" simulation of reionization (see Figure 9.8 and associated discussion for more information). In each panel, we set (by hand) the mean free path of ionizing photons to 10 comoving Mpc; the four columns show different H II fractions at a fixed redshift $z = 10$. The upper panels show the ionizing flux (arbitrary normalization) inside ionized zones according to the color scale. The lower panels show the locations of halos (dark points inside the white regions) and absorbers (green points). Crociani, D., et al., *Mon. Not. R. Astron. Soc.* **411**, 289 (2011). Copyright 2011 by the Royal Astronomical Society.



portion / PE

Plate 18 A ~~propotion~~ of the first Hubble Deep Field (HDF) image taken in 1995 (see Figure 10.1 and associated discussion for more information). The HDF covers an area 2.5 arcmin across and contains a few thousand galaxies (with a few candidates up to a $z \sim 6$). The image was taken in four broadband filters centered on wavelengths of 3000, 4500, 6060, and 8140Å, with an average exposure time of $\sim 1.27 \times 10^5$ s million seconds per filter.

insert space / PE

Plate 19 A full-scale model of the James Webb Space Telescope (JWST), the successor to the Hubble Space Telescope (http://www.jwst.nasa.gov/; see Figure 10.2 and associated discussion for more information). JWST includes a primary mirror 6.5 m in diameter and offers instrument sensitivity across the IR wavelength range of 0.6 to 28 μm, which will allow detection of the first generations of galaxies. The size of the Sun shield (the large flat screen in the image) is 22 in $\times$ 10 m (72 ft $\times$ 29 ft). The telescope will orbit 1.5 million km from Earth at the Lagrange L2 point. Courtesy of / NASA/EPO.

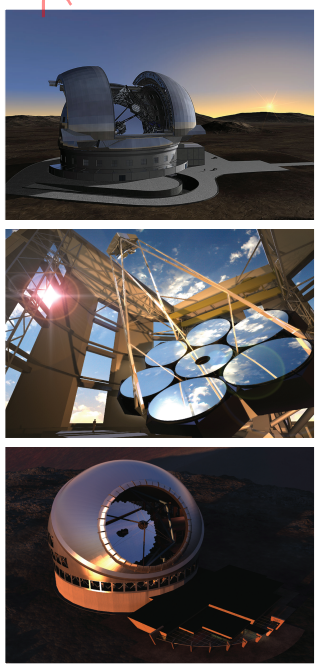


Plate 20 Artist's conception of the designs for three future giant ground-based telescopes that will be able to probe the first generation of galaxies (see Figure 10.3 and associated discussion for more information): the European Extremely Large Telescope (EELT, top), the Giant Magellan Telescope (GMT, middle), and the Thirty Meter Telescope (TMT, bottom). The European Southern Observatory (ESO), the GMT Partnership, and the TMT Observatory Corporation.

Courtesy of the

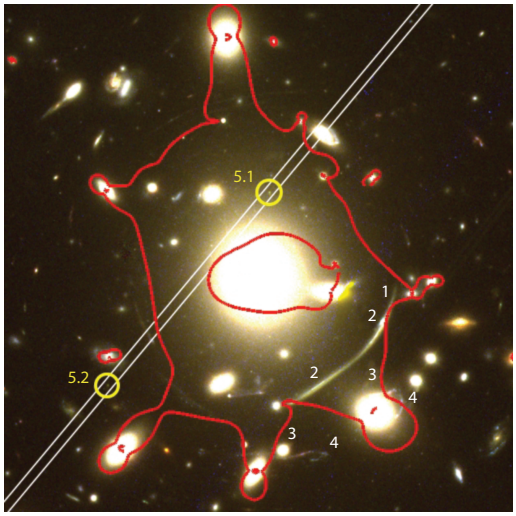


Plate 21 Image of the core of the cluster Abell 383, including known multiply imaged sources (marked 1 to 5) and the noncircular critical lines for sources at $z_s = 6$ (see Figure 10.7 and associated discussion for more information). The two images of a galaxy with $z_s = 6.027$ are marked by circles. The long slit used for spectroscopic follow-up is shown in white. Richard, J., et al., *Mon. Not. R. Astron. Soc.* **414**, L31 (2011). Copyright 2011 by the Royal Astronomical Society.

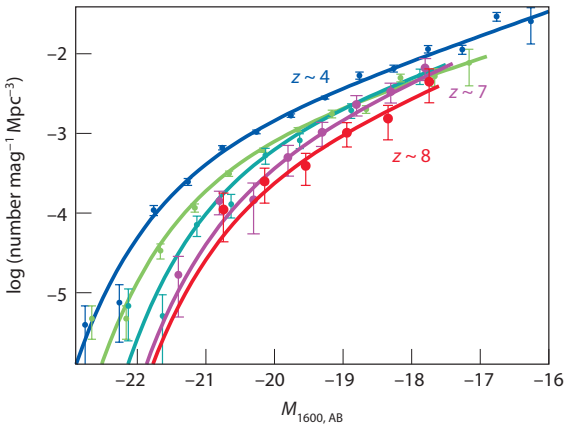
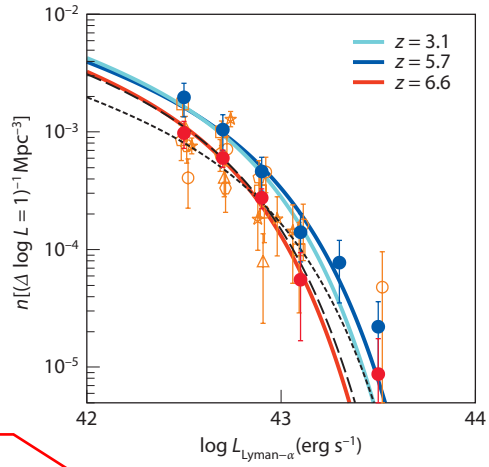


Plate 22 Rest-frame UV luminosity functions derived for galaxies at $z \sim 7$ (large filled circles) and $z \sim 8$ (large open circles) compared with lower-redshift data (from $z = 6$ to $z = 4$; the density increases as redshift decreases; see Figure 10.9 and associated discussion for more information). The vertical axis gives the number of galaxies per comoving cubic megaparsecs per AB magnitude at a rest-frame wavelength of 1600 Å, as a function of this magnitude on the horizontal axis. Note the sharp decline in the number density of bright galaxies with redshift and tentative evidence for a steepening faint-end slope. Bouwens, R., et al., *Astrophys. J.* **737**, 90 (2011). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by /
AAS. /



blue / AA

Plate 23 Luminosity function of LAEs for the range $z = 3$–6.6 (see Figure 10.11 and associated discussion for more information). The light solid circles show the measured luminosity function at $z = 6.6$, while the darker solid circles show the same for $z = 5.7$. The solid lines show Schechter function fits to these as well as the best fit at $z = 3.1$ (lightest curve). The LAE density drops substantially from $z = 5.7$ to $z = 6.6$, much faster than that of LBGs, but is nearly constant at lower redshifts. Finally, the open symbols show the number densities measured in the five subfields of the $z = 6.6$ survey, illustrating the substantial variance between fields. Ouchi, M., et al., *Astrophys. J.* **723**, 869 (2010). Reproduced with permission of the American Astronomical Society.

by / AAS. /

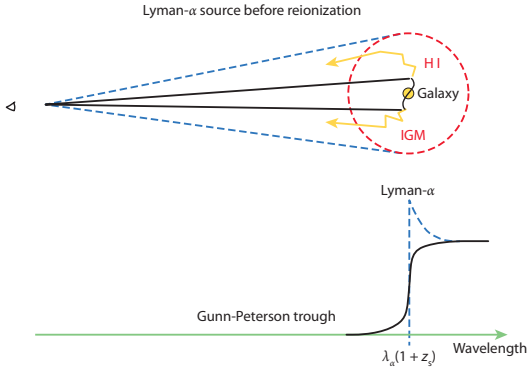


Plate 24 Halo of scattered Lyman-α line photons from a galaxy embedded in the neutral IGM prior to reionization (also called a *Loeb-Rybicki halo*; see Figure 11.10 and associated discussion for more information). The line photons diffuse in frequency owing the Hubble expansion of the surrounding medium 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 $\sim$1000 km s^{-1} relative to other lines (such as Hα) emitted by the galaxy.

to / PE

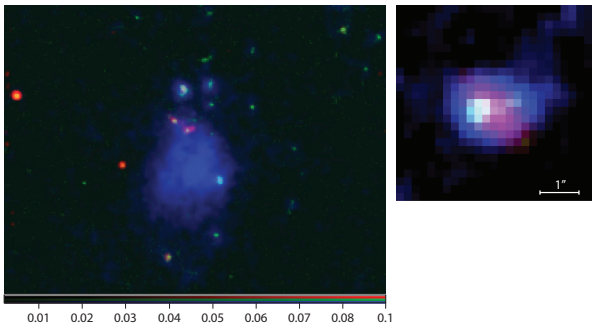


Plate 25 *Left:* A false-color image of a Lyman-α blob (LAB) at redshift $z = 2.656$. The hydrogen Lyman-α emission is shown in blue, and images in the optical V-band and the near-IR J and H bands are shown in green and red, respectively. Note the compact galaxies lying near the northern (top) end of 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. Prescott, M., & Dey, A. (2010). *Right:* A false-color image of a LAB at redshift $z = 6.6$, obtained from a combination of images at different IR wavelengths. (see Figure 11.12 and associated discussion for more information) Ouchi, M., et al., *Astrophys. J.* **696**, 1164 (2009). Reproduced ~~with~~ permission of the ~~American Astronomical Society.~~

by / AAS. /

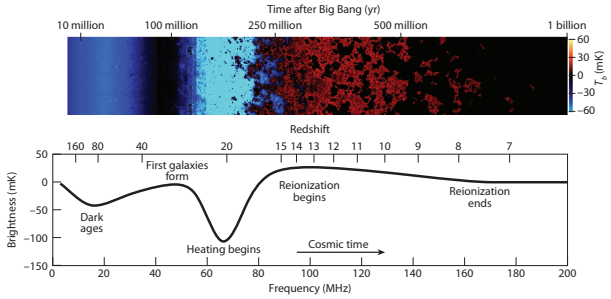


Plate 26 Overview of the expected 21-cm signal (see Figure 12.2 and associated discussion for more information). *Top panel:* Time evolution of fluctuations in the 21-cm brightness from just before the first stars form to the end of reionization. This evolution is pieced together from instantaneous redshift slices through a $(100\ \mathrm{Mpc})^3$ numerical simulation volume. Coloration indicates the strength of the 21-cm brightness as it transitions from absorption (blue) to emission (red) and finally disappears (black) owing to ionization. *Bottom panel:* Expected evolution of the sky-averaged 21-cm brightness from the dark ages at $z = 150$ to the end of reionization sometime before $z = 6$. The frequency structure is driven by the interplay of gas heating, the coupling of gas and 21-cm temperatures, and the ionization of the gas. The considerable uncertainty in the exact form of this signal arises from the poorly understood properties of the first galaxies. Pritchard, J. R., & Loeb, A., *Nature* **468**, 772 (2010). Copyright 2010 by Nature Publishing Group.

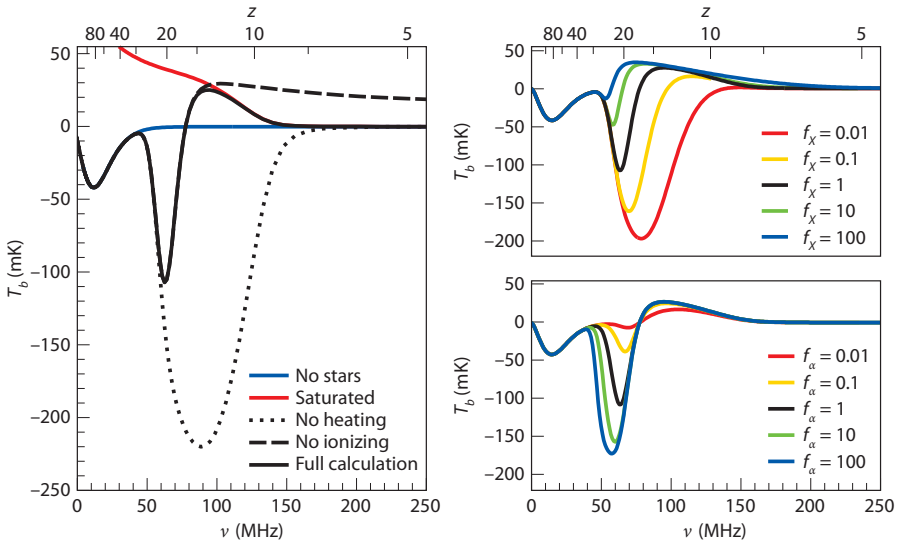


Plate 27 Monopole of the spin-flip brightness temperature (the so-called global 21-cm signal) in several models of early structure formation (see Figure 12.9 and associated discussion for more information). *Left:* Major variations around our fiducial model (solid curve with several turning points), as indicated. Each curve either eliminates a physical process (like heating or ionization) or maximizes it. *Right:* Suites of models in which the Lyman-α (lower panel) and X-ray heating (upper panel) efficiencies are varied by a factor of 10^4. Pritchard, J. R., & Loeb, A., *Phys. Rev.* **D82**, 023006 (2010). Copyright 2010 by the American Physical Society.

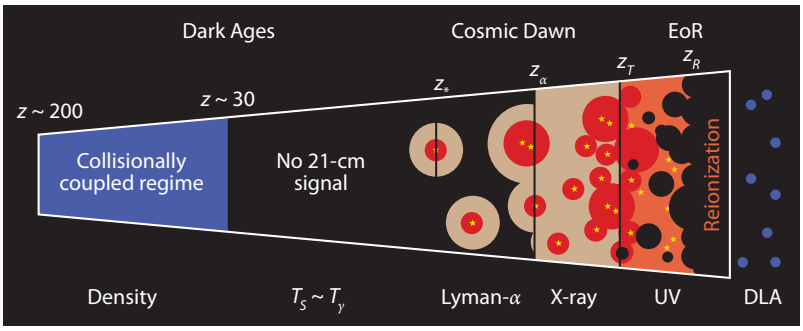


Plate 28 Cartoon of the different phases of the 21-cm signal (see Figure 12.10 and associated discussion for more information). The signal transitions from an early phase of collisional coupling to a later phase of Lyman-α coupling through a short period where there is little signal. Fluctuations after this phase are dominated successively by spatial variation in the Lyman-α, X-ray, and ionizing UV radiation backgrounds. After reionization is complete, there is a residual signal from neutral hydrogen in galaxies. Pritchard, J. R., & Loeb A., *Reports on Progress in Physics*, in press (2012), http://arxiv.org/pdf/1109.6012.

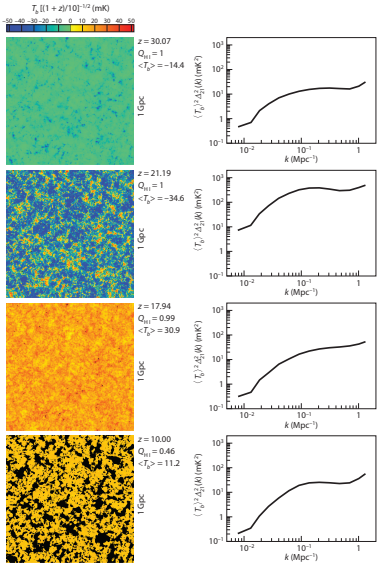


Plate 29 Slices through a "seminumerical" simulation (*left*), and the corresponding spherically averaged power spectra (*right*), for a model of the spin-flip background at $z = 30.1, 21.2, 17.9, 10.0$ (*top to bottom*; see Figure 12.14 and associated discussion for more information). The slices were chosen to highlight various epochs in the cosmic 21-cm signal (from top to bottom): the onset of Lyman-α pumping (here the blue regions show the cold gas around the first galaxies); the onset of X-ray heating (here the blue regions are cold gas, while the compact red regions represent hot gas around the first black holes); the completion of X-ray heating (where all the gas is hot); and the midpoint of reionization (where black regions are ionized bubbles). All slices are 1 Gpc on a side and 3.3 Mpc deep. Mesinger, A., Furlanetto, S. R., & Cen, R., *Mon. Not. R. Astron. Soc.* **411**, 955 (2011). Copyright 2011 by the Royal Astronomical Society.

Plate 30 One of the antenna "tiles" used in the Murchison Widefield Array (MWA) experiment in Western Australia (see Figure 12.19 and associated discussion for more information). Each such tile is composed of 16 crossed-dipole antennae, with their signals combined through hardware at the station. The full telescope combines the signals from ∼128 tiles interferometrically. This allows for a large (several hundred square degree) field of view with a moderately large collecting area. The antennas operate in the range 80–300 MHz, corresponding to $z \approx 6$–15 (although the telescope will be sensitive to the spin-flip background only at $z < 10$). C. Lonsdale.

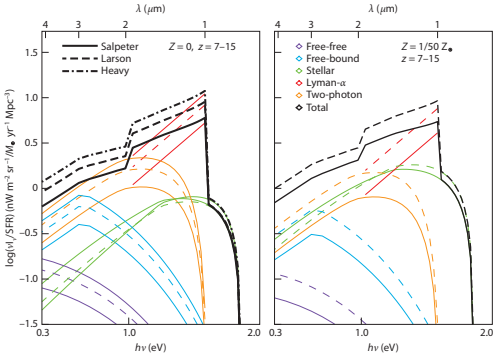


Plate 31 Near-IR background spectrum (see Figure 13.6 and associated discussion for more information). This example includes star formation ~~from~~ in the range $z = 7$–15, which dominates the background between 1 and 2 μm. The left panel assumes metal-free stars, while the right panel takes a metallicity $Z = 0.02\ Z_\odot$. The different line styles show different IMFs (see text). The different curves show the total emission, the Lyman-α contribution, the stellar continua, the two-photon continua, the free-bound emission, and finally the free-free emission (from top to bottom at 1 μm). Fernandez, E. R., & Komatsu, E., *Astrophys. J.* **646**, 703 (2006). Reproduced ~~with~~ permission of the ~~American Astronomical Society~~.

PE

by / AAS. /

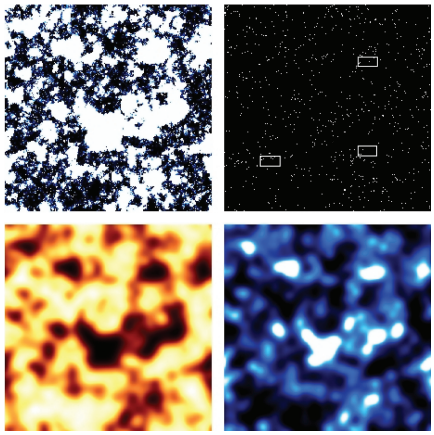


Plate 32 Schematic illustration of the cross-correlation between the spin-flip background and galaxy probes (see Figure 13.10 and associated discussion for more information). Each panel shows a slice from a numerical simulation ~185 Mpc across. *Upper left:* Map of the ionized fraction midway through reionization (ionized regions are shown in white). *Upper right:* Map of the galaxy field. White squares show regions $\sim 4 \times 8$ arcmin across; these are each equivalent to *four* adjacent pointings of the James Webb Space Telescope. *Lower left:* Spin-flip background from this slice. *Lower right:* Galaxy map smoothed over a 6-arcmin beam, as might be observed in an intensity mapping measurement. Note the strong large-scale anticorrelation between the bottom two figures. A. Lidz.

Courtesy of / AA