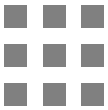
# MODERN COSMOLOGY

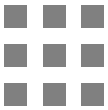
Second Edition

Scott Dodelson
Fabian Schmidt





# Modern Cosmology



# Modern Cosmology

## Second Edition

Scott Dodelson
Carnegie Mellon University
Department of Physics
Pittsburgh, PA, USA

Fabian Schmidt
Max-Planck-Institut für Astrophysik
Garching, Germany





ACADEMIC PRESS
An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2021 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

**Notices**

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

**Library of Congress Cataloging-in-Publication Data**
A catalog record for this book is available from the Library of Congress

**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from the British Library

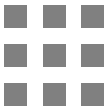
ISBN: 978-0-12-815948-4

For information on all Academic Press publications
visit our website at https://www.elsevier.com/books-and-journals

*Publisher:* Katey Birtcher
*Editorial Project Manager:* Susan Ikeda
*Production Project Manager:* Beula Christopher
*Designer:* Bridget Hoette

Typeset by VTeX





# Contents


About the authors xiii

Preface xv

1. The concordance model of cosmology 1
   - 1.1. A nutshell history of the universe 1
   - 1.2. The Hubble diagram 6
   - 1.3. Big Bang nucleosynthesis 8
   - 1.4. The cosmic microwave background 9
   - 1.5. Structure in the universe 10
   - 1.6. ΛCDM: the concordance model of cosmology 15
   - 1.7. Summary and outlook 16
   - Exercises 17

2. The expanding universe 21
   - 2.1. Expanding space 21
     - 2.1.1. The metric 22
     - 2.1.2. The geodesic equation 26
   - 2.2. Distances 30
   - 2.3. Evolution of energy 34
   - 2.4. Cosmic inventory 40
     - 2.4.1. Photons 40
     - 2.4.2. Baryons 41
     - 2.4.3. Dark matter 42
     - 2.4.4. Neutrinos 43
     - 2.4.5. Epoch of matter–radiation equality 47

2.4.6. Dark energy 47
2.5. Summary 52
Exercises 53

3. The fundamental equations of cosmology 57
3.1. Einstein equations 57
3.2. Boltzmann equation 62
3.2.1. Boltzmann equation for particles in a harmonic potential 63
3.2.2. Boltzmann equation in an expanding universe 65
3.2.3. Collision terms 68
3.3. Beyond the homogeneous universe 70
3.3.1. Perturbed spacetime 71
3.3.2. The geodesic equation 73
3.3.3. The collisionless Boltzmann equation for radiation 77
3.3.4. The collisionless Boltzmann equation for massive particles 78
3.4. Summary 79
Exercises 81

4. The origin of species 85
4.1. The homogeneous Boltzmann equation revisited 85
4.2. Big Bang nucleosynthesis 88
4.2.1. Neutron abundance 91
4.2.2. Light element abundances 94
4.3. Recombination 95
4.4. Dark matter 99
4.5. Summary 106
Exercises 107

5. The inhomogeneous universe: matter & radiation 111

5.1. The collisionless Boltzmann equation for photons 112
5.2. Collision terms: Compton scattering 114
5.3. The Boltzmann equation for photons 119
5.4. The Boltzmann equation for cold dark matter 122
5.5. The Boltzmann equation for baryons 126
5.6. The Boltzmann equation for neutrinos 129
5.7. Summary 130
Exercises 133

6. The inhomogeneous universe: gravity 135
6.1. Scalar–vector–tensor decomposition 135
6.2. From gauge to gauge 137
6.3. The Einstein equations for scalar perturbations 141
6.3.1. Ricci tensor 141
6.3.2. Two components of the Einstein equations 143
6.4. Tensor perturbations 147
6.4.1. Christoffel symbol for tensor perturbations 148
6.4.2. Ricci tensor for tensor perturbations 149
6.4.3. Einstein equations for tensor perturbations 150
6.4.4. Verifying the decomposition theorem 153
6.5. Summary 154
Exercises 155

7. Initial conditions 157
7.1. The horizon problem and a solution 157
7.2. Inflation 163
7.3. Gravitational wave production 167
7.3.1. Quantizing the harmonic oscillator 168
7.3.2. Tensor perturbations 170
7.4. Scalar perturbations 173

7.4.1. Scalar field perturbations around an unperturbed background 175
7.4.2. Super-horizon perturbations 177
7.4.3. Spatially flat slicing 181
7.5. The Einstein–Boltzmann equations at early times 183
7.6. Summary 186
Exercises 189

8. Growth of structure: linear theory 195
8.1. Prelude 195
8.1.1. Three stages of evolution 196
8.1.2. Closing the Boltzmann hierarchy 200
8.2. Large scales 203
8.2.1. Super-horizon solution 203
8.2.2. Through horizon crossing 207
8.3. Small scales 209
8.3.1. Horizon crossing 209
8.3.2. Sub-horizon evolution 214
8.4. The transfer function 217
8.5. The growth factor 220
8.6. Beyond cold dark matter and radiation 222
8.6.1. Baryons 222
8.6.2. Massive neutrinos 224
8.6.3. Dark energy 225
8.7. Summary 225
Exercises 226

9. The cosmic microwave background 231
9.1. Overview 231
9.2. Large-scale anisotropies 237

9.3. Acoustic oscillations 238
9.3.1. Tightly-coupled limit of the Boltzmann equations 238
9.3.2. Tightly-coupled solutions 242
9.4. Diffusion damping 244
9.5. Inhomogeneities to anisotropies 247
9.5.1. Free streaming 247
9.5.2. The angular power spectrum 251
9.6. The CMB power spectrum 255
9.6.1. Large angular scales 255
9.6.2. Acoustic peaks 257
9.7. Cosmological parameters 259
9.7.1. Curvature and $\Lambda$ 260
9.7.2. Amplitude, spectral index, and optical depth 262
9.7.3. Baryon and CDM densities 263
9.8. Summary 265
Exercises 266

10. The polarized CMB 271
10.1. Polarization 271
10.2. Generating polarization from Compton scattering 275
10.3. Polarization from a single plane wave 278
10.4. Boltzmann solution 283
10.5. Polarization power spectra 285
10.6. Detecting gravitational waves 288
10.7. Summary 291
Exercises 292

11. Probes of structure: tracers 295
11.1. Galaxy clustering 296
11.1.1. Galaxy statistics 299

11.1.2. Redshift-space distortions 301
11.1.3. BAO and Alcock–Paczyński 305
11.2. Angular correlations 311
11.3. The Sunyaev–Zel'dovich effect 315
11.4. Summary 320
Exercises 321

12. Growth of structure: beyond linear theory 325
12.1. Prelude 326
12.2. Perturbation theory 330
12.3. Simulations 343
12.4. Dark matter halos 347
12.4.1. Halo masses and profiles 349
12.4.2. The halo mass function 350
12.5. Galaxy clusters 355
12.6. Galaxy clustering and bias 359
12.7. The halo model 364
12.8. Summary 366
Exercises 369

13. Probes of structure: lensing 373
13.1. Overview 373
13.2. Photon geodesics 376
13.3. CMB lensing 381
13.4. Galaxy shapes 384
13.5. Weak-lensing statistics 388
13.5.1. Shear power spectrum 388
13.5.2. Shear correlation function 391
13.5.3. Shear cross-correlations 393
13.6. Summary 396

Exercises 397

14. Analysis and inference 401
14.1. The likelihood function 402
14.2. Overview: from raw data to parameter constraints 407
14.3. Mapmaking 409
14.4. Two-point functions 412
14.4.1. CMB power spectrum 413
14.4.2. Galaxy power spectrum 417
14.5. The Fisher matrix 421
14.6. Sampling the likelihood function 424
14.7. Summary 427
Exercises 428

A. Solutions to selected exercises 433

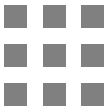
B. Numbers 465
B.1. Physical constants 465
B.2. Astrophysical constants 465
B.3. Fiducial cosmology 466

C. Special functions 467
C.1. Legendre polynomials 467
C.2. Spherical harmonics 467
C.3. Spherical Bessel functions 469
C.4. Fourier transforms 470
C.5. Miscellaneous 470

D. Symbols 473
D.1. Mathematical and geometrical definitions 473
D.2. Frequently used relations 473
D.3. Symbol definitions 474

Bibliography 477

Index 483



# About the authors

**Scott Dodelson** is Head of the Department of Physics at Carnegie Mellon University. He received his Ph.D. from Columbia University and was a research fellow at Harvard before moving to Fermilab, the University of Chicago, and Carnegie Mellon. He is the author of more than 200 papers on cosmology, most of which focus on the cosmic microwave background and the large-scale structure of the universe. Dodelson serves as the co-Chair of the Science Committee of the Dark Energy Survey.

**Fabian Schmidt** is a research group leader at the Max-Planck-Institut für Astrophysik (MPA) in Garching, Germany. He obtained his Ph.D. in Astronomy & Astrophysics at the University of Chicago in 2009, and moved to MPA after fellowships at Caltech and Princeton. His work in cosmology (approximately 100 papers so far) focuses on theory, numerics, and analysis of quasilinear and nonlinear large-scale structure, and how we can learn from it about gravity, dark energy, and the physics of inflation.

# Preface

Since the first edition of *Modern Cosmology* was published in 2003, cosmology has made dramatic advances, both theoretical and experimental. We thought (as perhaps many others) that it would be useful to revise the first edition to reflect these developments. On the one hand, the changes in the cosmological landscape are significant: the old Euclidean $\Omega_{\rm m} = 1$ ("sCDM") model was on its last legs in 2003 (though it dominated many of the plots in the first edition) and has now passed on; measurements of polarization and weak lensing were in their infancy; optimism about WIMP detection was high; Markov Chain Monte Carlo sampling was not yet in widespread use; and baryon acoustic oscillations had not been detected (one of us apparently thought they were "barely (if at all) detectable"). On the other hand, the basics have not changed much: the same equations still govern the evolution of the universe; the core paradigm of cosmology is still that large-scale structure in the universe emerged from gravitational collapse of small perturbations generated very early in time; the CMB and large-scale structure are still thought to be the primary way of learning about cosmology; and analysis is more important than ever.

This new edition improves on the old not just by updating plots to reflect the more recent results, but also by restructuring and filtering the old material so that it now satisfies not just one author but two. A total of three new chapters have been added. The most significant addition in terms of content is nonlinear structure formation (Ch. 12) which has become a major focus of cosmology over the past decade; but the new material on the above-mentioned baryon acoustic oscillations, the Sunyaev–Zel'dovich effect, CMB lensing, as well as Markov Chain Monte Carlo is also worth highlighting.

We are grateful to many colleagues for their contributions to this second edition. Michael Blanton kindly provided glorious pictures of large-scale structure. Julien Lesgourgues was of crucial help in producing the plots of results of the CLASS code in Chs. 8–9. Giovanni Cabass helped make Fig. 9.4. Florian Beutler provided Fig. 11.7, and Lindsey Bleem supplied us with Fig. 12.10.

We are further indebted to Elisabeth Krause for feedback that helped improve many chapters, but Ch. 8 and Ch. 14 in particular. We also thank Vincent Desjacques and Donghui Jeong for their comments on Ch. 12. FS thanks the members of the 2019 Garching book club—Alex Barreira, Philipp Busch, Chris Byrohl, Giovanni Cabass, Dani Chao, Daniel Farrow, Laura Herold, Jiamin Hou, Martha Lippich, Kaloian Lozanov, Leila Mirzagholi, Minh Nguyen, Yuki Watanabe, and Sam Young—for their innumerable comments and suggestions, as well as for pointing out many errors. FS is also very grateful to the Junge Akademie for enabling several writing retreats. SD thanks Nianyi Chen, Biprateep Dey, and Kuldeep Sharma for getting the hint right in Exercise 1.2, and Troy Raen for comments on Ch. 2.

Many thanks go to Robert Smith, for pointing out errors in Exercise 7.7, to Tom Crawford for answering our questions on mapmaking algorithms used in current ground-based CMB analyses, and to Eiichiro Komatsu for helpful discussions on how to calculate the energy-momentum tensor. We are grateful to Giovanni Cabass and Yuki Watanabe for help getting the factors of 2 in the Compton collision term correct.

Further, we thank all readers of the first edition who took time to report typos and mistakes therein which, where still relevant, were corrected here. Undoubtedly, errors are also lurking in this edition; any corrections and feedback are most welcome under modcosmology@gmail.com.

Finally, FS thanks SD for making this second edition possible (which generally requires a first edition), and for making an offer he could not refuse. SD is extremely grateful that the person best suited to write this book agreed to be a co-author.

Scott Dodelson
Pittsburgh, PA, United States
Fabian Schmidt
Garching, Germany
October 1, 2019

# 1

# The concordance model of cosmology

Einstein's discovery of general relativity in the previous century enabled us, for the first time in history, to come up with a compelling, testable theory of the universe. The realization that the universe is expanding and was once much hotter and denser allows us to modernize the deep age-old questions "Why are we here?" and "How did we get here?" The updated versions are now "How did the elements form?", "Why is the universe so smooth?", and "How did galaxies form within this smooth universe?" Remarkably, these questions and many like them have quantitative answers, answers that can be found only by combining our knowledge of fundamental physics with our understanding of the conditions in the early universe. Even more remarkably, these answers can be tested against astronomical observations. Before going into depth, we begin with a broad-brush overview of our current state of knowledge on the history of the universe in this chapter and the next.

The success of the Big Bang paradigm rests on a number of observational pillars: the Hubble diagram that measures expansion; light element abundances that are in accord with Big Bang Nucleosynthesis; temperature and polarization anisotropies in the cosmic microwave background that agree well with theory; and multiple probes of large-scale structure that also agree with the concordance model that will be described in this Chapter. This success has come at a price, however: we are now forced to introduce several ingredients that go beyond the Standard Model of particle physics (for a quick overview, see Box 1.1):

- dark matter and dark energy, which together dominate the energy budget of the universe over most of its lifetime;
- a mechanism generating the small initial perturbations out of which structure formed, the most popular explanation being inflation.

## 1.1 A nutshell history of the universe

We have solid evidence that the universe is expanding. This means that early in its history the distance between us and distant galaxies was smaller than it is today. It is convenient to describe this effect by introducing the scale factor $a$, whose present value is set to 1 by convention. At earlier times, $a$ was smaller than it is today. We can imagine placing a grid in space as in Fig. 1.1 which expands uniformly as time evolves. Points on the grid, which correspond to observers at rest, maintain their coordinates, so the *comoving distance* between two points—which just measures the difference between coordinates, and can be obtained

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00007-3
Copyright © 2021 Elsevier Inc. All rights reserved.

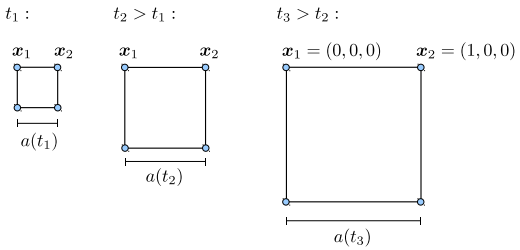



**FIGURE 1.1** Expansion of the universe. The comoving distance between points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ on a grid that serves as a coordinate system remains constant as the universe expands; in this case, $|\boldsymbol{x}_2 - \boldsymbol{x}_1| = 1$. The physical distance is proportional to the comoving distance times the scale factor, so it grows as time evolves.

by counting grid cells as indicated in Fig. 1.1—remains constant. However, the physical distance is proportional to the scale factor, and the physical distance does evolve with time.

A directly related effect is that the physical wavelength of light emitted from a distant object is stretched out proportionally to the scale factor, so that the observed wavelength is larger than the one at which the light was emitted. It is convenient to define this stretching factor as the redshift $z$:

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} = \frac{a_{\text{obs}}}{a_{\text{emit}}} = \frac{1}{a_{\text{emit}}}. \tag{1.1}$$

In addition to the scale factor and its evolution, the smooth universe is characterized by one other parameter, its geometry. There are three possibilities: Euclidean, open, or closed universes. These different possibilities are best understood by considering two freely traveling particles which start their journeys moving parallel to each other. In a *Euclidean* universe, often also called a "flat universe," the particles behave as Euclid himself expected them to: their trajectories remain parallel as long as they travel freely. If the universe is *closed*, the initially parallel particles gradually converge, just as in the case of the 2-sphere all lines of constant longitude meet at the North and South Poles. The analogy of a closed universe to the surface of a sphere runs even deeper: both are spaces of *constant positive curvature*, the former in three spatial dimensions and the latter in two. Finally, in an *open* universe, the initially parallel paths diverge, as would two marbles rolling off a saddle.

General relativity connects geometry to energy. Accordingly, the total energy density in the universe determines the geometry: if the density is higher than a critical value, $\rho_{\text{cr}}$, which we will soon see is approximately $10^{-29}\,\text{g}\,\text{cm}^{-3}$, the universe is closed; if the density is lower, it is open. A Euclidean universe is one in which the energy density is precisely equal to critical. This seems unlikely to happen, and yet all observations indicate that the universe is Euclidean to within errors! We will later see that inflation provides a natural explanation for this fact.

To understand the history of the universe, we must determine the evolution of the scale factor $a$ with cosmic time $t$. Again, general relativity provides the connection between this evolution and the energy in the universe. Fig. 1.2 shows how the scale factor increases as

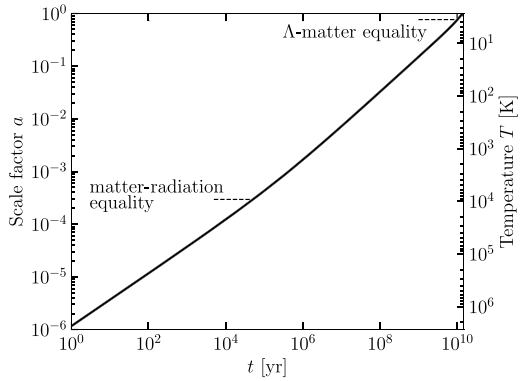



**FIGURE 1.2** Evolution of the scale factor of the universe with cosmic time. The universe today corresponds to the upper-right corner of the plot, where $a(t_0) = 1$ and the temperature $T = 2.73$ K. When the universe was very young, radiation was the dominant component, and the scale factor increased as $t^{1/2}$. At the indicated point, the universe transitioned to matter domination, during which $a(t) \propto t^{2/3}$. Very recently, the expansion law changed again due to dark energy, with $a(t)$ transitioning to an exponential function of time.

the universe ages, with the convention that $a = 1$ today. Note that the dependence of $a$ on $t$ varies as the universe evolves. At early times, $a \propto t^{1/2}$ while at later times the dependence switches to $a \propto t^{2/3}$. How the scale factor varies with time is determined by the evolution of the energy density in the universe. At early times, radiation dominates, while at later times, nonrelativistic matter accounts for most of the energy density. In fact, one way to explore the energy content of the universe is to measure changes in the scale factor. We will see that, partly as a result of such an exploration, we now know that $a$ has been growing faster than $t^{2/3}$ very recently, a signal that a new form of energy has come to dominate the late-time cosmological landscape.

To quantify the change in the scale factor and its relation to the energy, it is useful to define the Hubble rate

$$H(t) \equiv \frac{1}{a}\frac{da}{dt}, \tag{1.2}$$

which measures how rapidly the scale factor changes. For example, if the universe is Euclidean and matter dominated, so that $a \propto t^{2/3}$, then $H = (2/3)t^{-1}$. Throughout this book, a subscript 0 will denote the value of a quantity today. $H_0 \equiv H(t_0)$ is known as *Hubble's constant*. Thus, in a Euclidean, matter-dominated universe (not ours!), the product $H_0 t_0$ equals 2/3.

More generally, general relativity predicts that the scale factor is determined by the Friedmann equation (which we will derive in Ch. 3):

$$H^2(t) = \frac{8\pi G}{3}\left[\rho(t) + \frac{\rho_{\rm cr} - \rho(t_0)}{a^2(t)}\right] \tag{1.3}$$

where $G$ is Newton's constant; $\rho(t)$ is the energy density in the universe as a function of time with $\rho(t_0)$ its value today; and $\rho_{\rm cr}$ is the aforementioned *critical density*. It is a constant

given by

$$\rho_{\rm cr} \equiv \frac{3H_0^2}{8\pi G}. \tag{1.4}$$

Eq. (1.3) allows for the possibility that the universe is not Euclidean: if it were Euclidean, the sum of all the energy densities today would equal the critical density, and the last term in Eq. (1.3) would vanish. If the universe is not Euclidean, the curvature contribution scales as $1/a^2$. In most of this book we will work within the context of a Euclidean universe since there are several persuasive arguments, both observational and theoretical, that support this assumption. We will learn about these arguments in Ch. 2 and Ch. 7.

To use the Friedmann equation, we must know how the energy density evolves with time. This turns out to be a complicated question because $\rho$ in Eq. (1.3) is the sum of several different components, each of which scales differently with time. Consider first nonrelativistic matter, which means that the energy of a given constituent particle is essentially equal to its rest mass energy, which remains constant with time. The energy density of a collection of these particles is therefore equal to the rest mass energy times the number density. When the scale factor was smaller, the densities were necessarily larger. Since number density is inversely proportional to volume, it should be proportional to $a^{-3}$. Therefore the energy density of nonrelativistic matter scales as $a^{-3}$.

Apart from matter, a sea of massless photons permeates the universe, as first discovered in 1965. These photons have traveled freely since the universe was very young. Today, their wavelengths lie in the microwave regime, so they comprise what is called the cosmic microwave background (CMB). The CMB has a perfect black-body spectrum with a very well-measured temperature of $T_0 = 2.726 \pm 0.001$ K today (Fixsen, 2009). Our redshift relation Eq. (1.1) allows us to derive how this temperature evolved over the history of the universe. Since $\lambda = c/\nu \propto a$, the frequency $\nu$ of any photon decays as $1/a$ with the expansion. The black-body spectrum is a function of $\nu/T$, so we can describe this effect equivalently by stating that the temperature of the radiation as a function of time is given by

$$T(t) = \frac{T_0}{a(t)}. \tag{1.5}$$

In the next chapter, we will rederive this result in a complementary way. At early times, then, the temperature was higher than it is today. The energy density of black-body radiation scales as $T^4 \propto a^{-4}$, as indicated in Fig. 1.3. Via Eq. (1.3), this implies that the Hubble parameter at early times evolves as $H \propto T^2$.

Fig. 1.3 illustrates how the different contributions to $\rho(t)$ in Eq. (1.3) vary with the scale factor. Early on, because of the $a^{-4}$ scaling, radiation was the dominant constituent of the universe, but today, matter and *dark energy*, which could be a *cosmological constant*, dominate the landscape. We will have more to say about dark energy later, but for now simply note that whether it does indeed contribute to the energy exactly as the constant depicted in Fig. 1.3 is still an open question.

Let us introduce some numbers. The expansion rate is a measure of how fast the universe is expanding, determined by measuring the velocities of distant galaxies and dividing

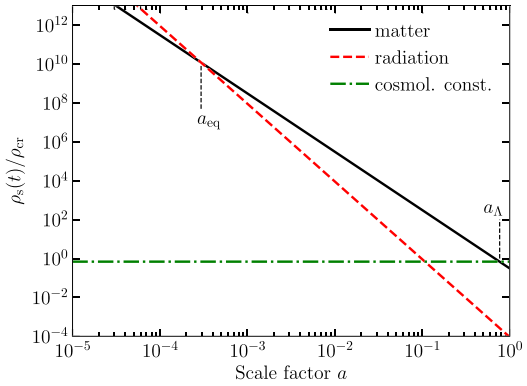



**FIGURE 1.3** Energy density as a function of scale factor for different constituents of the Euclidean fiducial cosmology, whose parameters are listed in Appendix B.3: nonrelativistic matter ($\propto a^{-3}$), radiation ($\propto a^{-4}$), and a cosmological constant. All are in units of the critical density today. Even though matter and the cosmological constant appear to dominate today, at early times, the radiation density was largest. The epoch at which the energy densities of matter and radiation are equal is $a_{\rm eq}$, while the epoch at which the densities of matter and cosmological constant match is $a_\Lambda$.

by their distance from us (Sect. 1.2). So the expansion is usually written in units of velocity per distance. The Hubble constant is parameterized by a dimensionless number $h$ defined via

$$\begin{aligned} H_0 &= 100\,h\ \mathrm{km\ s^{-1}\,Mpc^{-1}} \\ &= \frac{h}{0.98\times 10^{10}\ \mathrm{years}} = 2.13\times 10^{-33}\ \frac{\mathrm{eV}}{\hbar}\,h \end{aligned} \tag{1.6}$$

where $h$ has nothing to do with Planck's constant $\hbar$. The astronomical length scale of a megaparsec (Mpc) is equal to $3.0856\times 10^{24}$ cm. Current measurements yield $h\simeq 0.7$. Since Edwin Hubble's initial measurement in 1929, the value of the Hubble constant has been subject to vigorous debate, and even now there is some controversy about its precise value, at the 5% level. For this reason, it has become customary to use $h^{-1}$ Mpc as the unit of length in cosmology. With this unit, and some associated units such as $h^{-1}\,M_\odot$ for masses ($M_\odot$ denotes a solar mass), the Hubble constant drops out of many computations, so that they become insensitive to its precise value. We will follow this convention throughout the book as well.

The predicted age for a Euclidean, matter-dominated universe, $(2/3)H_0^{-1}$, is then $6.5h^{-1}$ Gyr. You will show in Exercise 1.2 that the age of a universe with a cosmological constant is larger (for fixed $h$). In fact one of the original arguments in favor of a cosmological constant was to make the universe older and thus compatible with the age estimates of the oldest observed stars (which are older than 10 billion years).

Newton's constant in Eq. (1.4) is equal to $6.67\times 10^{-8}\ \mathrm{cm^3\,g^{-1}\,s^{-2}}$. This, together with Eq. (1.6), enables us to get a numerical value for the critical density:

$$\rho_{\rm cr} = 1.88\,h^2\times 10^{-29}\ \mathrm{g\,cm^{-3}}. \tag{1.7}$$

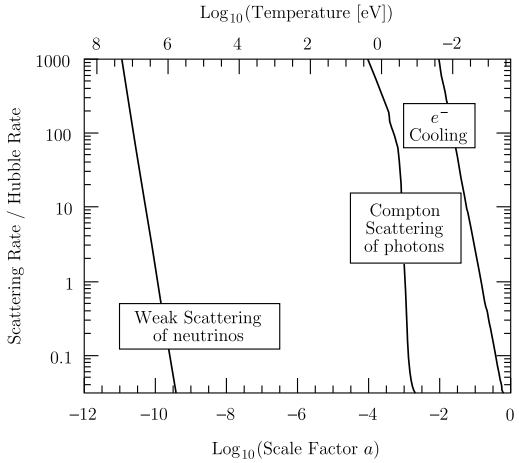



**FIGURE 1.4** Interaction rates as a function of the scale factor. When a given rate becomes smaller than the expansion rate $H$, that reaction falls out of equilibrium. Top scale gives ($k_B$ times) the temperature of the universe, an indication of the typical kinetic energy per particle.

An important ramification of the higher densities and temperatures in the past is that the rates for particles to interact with each other, which typically scale as the density squared, were also much higher early on. Fig. 1.4 shows some important rates as a function of the scale factor. For example, when the temperature of the universe was greater than several MeV/$k_B$, the rate for electrons and neutrinos to scatter was larger than the expansion rate. Thus, before the universe could double in size, a neutrino scattered many times off the ubiquitous electrons. All these scatterings brought the neutrinos into equilibrium with the rest of the cosmic plasma. This is but one example of a very general, profound fact: if a particle scatters with a rate much greater than the expansion rate, that particle stays in equilibrium. Otherwise, it falls out of equilibrium with the other species and "freezes out." Since rates were typically large, the early universe was a relatively simple environment: not only was it very smooth, but many of its constituents were in equilibrium. Ch. 2 explores some manifestations of the equilibrium conditions, while Ch. 4 touches on several cases where equilibrium could not be maintained because the reaction rates dropped beneath the expansion rate.

## 1.2 The Hubble diagram

If the universe is expanding as depicted in Fig. 1.1, then galaxies should be moving away from each other. We should therefore see galaxies receding from us. Hubble (1929) first found that distant galaxies are in fact all apparently receding from us, i.e. redshifted. He also noticed the trend that the velocity increases with distance. This is exactly what we expect in an expanding universe, for the physical distance between two galaxies is $d = ax$

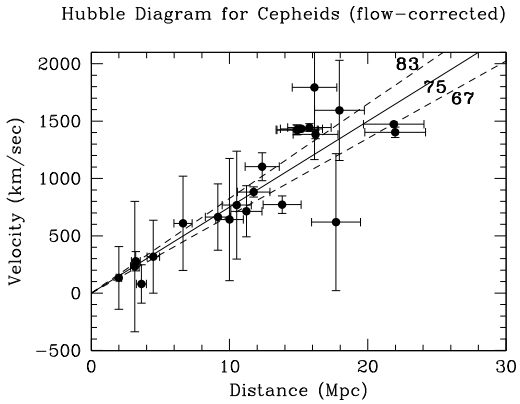



**FIGURE 1.5** A modern version of the Hubble diagram from the Hubble Space Telescope Key project (Freedman et al., 2001). Each point corresponds to a galaxy whose distance has been estimated using pulsating stars known as Cepheid variables. The recession velocity for each galaxy is then corrected using a model for the peculiar velocity field in the neighborhood of the Milky Way. The lines show the prediction of the Hubble-Lemaître law with different values of $H_0$ (in $\mathrm{km\,s^{-1}\,Mpc^{-1}}$), as indicated.

where $x$ is the comoving distance.[1] In the absence of any comoving motion, $\dot{x} \equiv dx/dt = 0$ (no *peculiar* velocity), the relative velocity $v$ is therefore equal to

$$v = \frac{d}{dt}(ax) = \dot{a}x = H_0\, d \qquad (v \ll c), \tag{1.8}$$

where overdots indicate derivatives with respect to time $t$. Therefore, the apparent velocity should increase linearly with distance (at least at low redshift) with a slope given by $H_0$, the Hubble constant. Eq. (1.8) is known as the *Hubble-Lemaître law*. The value of the constant is simply determined by measuring the slope of the line in the *Hubble diagram* shown in Fig. 1.5.

In the next chapter, we will generalize the distance-redshift relation to larger distances, where Eq. (1.8) breaks down. Instead of recession velocities, this more rigorous derivation will be based on the stretching of the wavelength of light encoded in Eq. (1.1). For now, let us just point out that the distance-redshift relation depends on the energy content of the universe. Data from a variety of sources point to a current best-fit scenario that is Euclidean and contains about 70% of the energy in the form of a cosmological constant, or some other form of dark energy. This now forms the concordance cosmology that will be our working model throughout.

[1] This is strictly true only for small values of $z$, or equivalently recession velocities much smaller than the speed of light $c$. We will discuss the subtleties associated with properly defining distances in an expanding universe in Sect. 2.2.

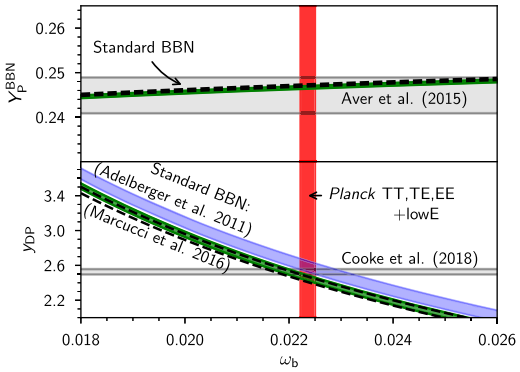



**FIGURE 1.6** Predicted primordial abundances (lines) of helium (top) and deuterium (bottom) as a function of the physical baryon density in units of $\rho_{cr}$, $\omega_b = \Omega_b h^2$. The subscript $_P$ on the $y$-axes denotes that these are the primordial abundances; $Y_P$ Is the ratio of the mass density in helium to the total mass density in protons and neutrons, while $y_D$ is defined as $10^5$ times the ratio of deuterium to hydrogen. The horizontal bands show astrophysical constraints on abundances, while the vertical band indicates the constraint based on CMB anisotropies, as measured by the Planck satellite experiment. In case of deuterium, the predictions are uncertain due to imperfect knowledge of certain nuclear reaction rates. Nevertheless, there is striking agreement between BBN (combined with astrophysical measurements) and the CMB. From Planck Collaboration (2018b).

## 1.3 Big Bang nucleosynthesis

Armed with an understanding of the evolution of the scale factor and the densities of the constituents in the universe, we can extrapolate backwards to explore phenomena at early times. When the universe was much hotter and denser, and the temperature was of order 1 MeV/$k_B$, there were no neutral atoms or even bound nuclei. The vast amounts of high-energy radiation in such a hot environment ensured that any atom or nucleus produced would be immediately destroyed by a high-energy photon. As the universe cooled well below typical nuclear binding energies, light elements began to form in a process known as *Big Bang Nucleosynthesis (BBN)*. Knowing the conditions of the early universe and the relevant nuclear cross-sections, we can calculate the expected primordial abundances of all the elements (Ch. 4).

Fig. 1.6 shows the BBN predictions for the abundances of helium and deuterium as a function of the mean *baryon density*, essentially the density of ordinary matter (Sect. 2.4) in the universe, in units of the critical density. The predicted abundances, in particular that of deuterium, which we will explore in detail in Ch. 4, depend on the density of protons and neutrons at the time of nucleosynthesis. The combined proton plus neutron density is equal to the baryon density since both protons and neutrons have baryon number one and these are the only baryons around at the time.

The horizontal lines in Fig. 1.6 show the current measurements of the light element abundances. The deuterium abundance is measured in the intergalactic medium at high redshifts by looking for a subtle absorption feature in the spectrum of distant quasars (see Burles and Tytler, 1998; Cooke et al., 2018 and Exercise 1.3). These measurements of the

abundances, combined with BBN calculations, give us a way of measuring the baryon density in the universe, constraining ordinary matter to contribute at most 5% of the critical density (note that the $x$-axis in Fig. 1.6 is the baryon density divided by the critical density, but multiplied by $h^2 \simeq 0.5$). Since the total matter density today is significantly larger than this—as we will see throughout the book—nucleosynthesis provides a compelling argument for matter that is comprised of neither protons or neutrons. This new type of matter has been dubbed *dark matter* because it apparently does not emit light. One of the central questions in physics now is: "What is the Dark Matter?"

## 1.4 The cosmic microwave background

Another phenomenon that falls out of energetics and a qualitative understanding of the evolution of the universe is the origin of the CMB. When the temperature of the radiation was of order $10^4$ K (corresponding to energies of order an eV), free electrons and protons combined to form neutral hydrogen. Before then, any hydrogen produced was quickly ionized by energetic photons. After that epoch, at $z \simeq 1100$, the photons that comprise the CMB ceased interacting with any particles and traveled freely through space. When we observe them today, we are thus looking at messengers from an early moment in the universe's history. They are therefore the most powerful probes of the early universe. We will spend an inordinate amount of time in this book working through the details of what happened to the photons before they last scattered off of free electrons, and also developing the mathematics of the free-streaming process since then. Among many other aspects, we will understand how the CMB constrains the baryon density independently, and in agreement with BBN as shown in Fig. 1.6, providing a ringing confirmation of the concordance model.

For now, we are only concerned with the crucial fact that the interactions of photons with electrons before last scattering ensured that the photons were in equilibrium. That is, they should have a black-body spectrum. The specific intensity of a gas of photons with a black-body spectrum is

$$I_\nu = \frac{4\pi\hbar\nu^3/c^2}{\exp[2\pi\hbar\nu/k_{\rm B}T] - 1}. \tag{1.9}$$

Fig. 1.7 shows the remarkable agreement between this prediction (see Exercise 1.4) of Big Bang cosmology and the observations by the FIRAS instrument aboard the COBE satellite. In fact, the CMB provides the best black-body spectrum ever measured. We have been told[2] that detection of the 3K background by Penzias and Wilson in the mid-1960s was sufficient evidence to decide the controversy in favor of the Big Bang over the Steady State universe, an alternative scenario without any expansion. Penzias and Wilson, though, measured the radiation at just one wavelength. If even their one-wavelength result was

[2] For a fascinating first-hand account of the history of the discovery of the CMB, see Ch. 1 of Partridge (2007).

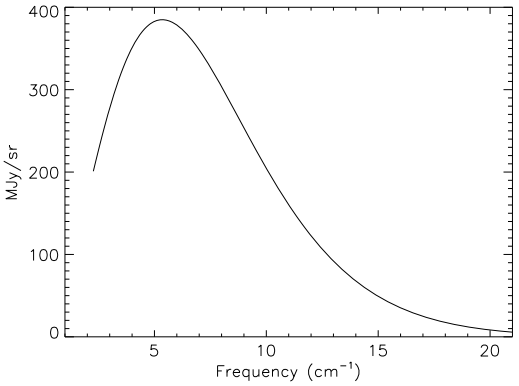



**FIGURE 1.7** Intensity of cosmic microwave radiation as a function of frequency from Far InfraRed Absolute Spectrophotometer (FIRAS), an instrument on the COBE satellite. The line shows a black-body spectrum with $T_0 = 2.728$ K. The error bars on the measurements are smaller than the width of the line! From Fixsen et al. (1996).

enough to tip the scales, the current data depicted in Fig. 1.7 should send skeptics from the pages of physics journals to the far reaches of radical internet chat groups.

The most important fact we learned from our first 25 years of surveying the CMB was that the early universe was very smooth. No anisotropies were detected in the CMB. This period, while undoubtedly frustrating for observers searching for anisotropies, solidified the view of a smooth Big Bang. The satellite mission COBE discovered anisotropies in the CMB in 1992, indicating that the early universe was not completely smooth. There were small perturbations in the cosmic plasma, with fractional temperature fluctuations of order $10^{-5}$. By now, these small fluctuations have been mapped with exquisite precision, and the state of the art is to look for even more subtle effects such as CMB polarization and the effect of the intervening matter distribution through gravitational lensing. To understand all of these effects, we must clearly go beyond the smooth background universe and look at deviations from smoothness, or *inhomogeneities*. Inhomogeneities in the universe are often simply called *structure*.

## 1.5 Structure in the universe

The existence of structure in the universe was known long before the detection of CMB anisotropies: various efforts to map out the distribution of galaxies in the local universe clearly showed that they are not distributed homogeneously. The number of galaxies and volume covered by such surveys has grown exponentially. Two surveys in particular broke new ground: the Sloan Digital Sky Survey (SDSS; Fig. 1.8) and the Two Degree Field Galaxy Redshift Survey (2dF), which between them compiled the redshifts of, and hence the distances to, over a million galaxies. Projects over the ensuing decades have and will provide deeper and more detailed maps than these ground-breaking surveys, by orders of magnitude.

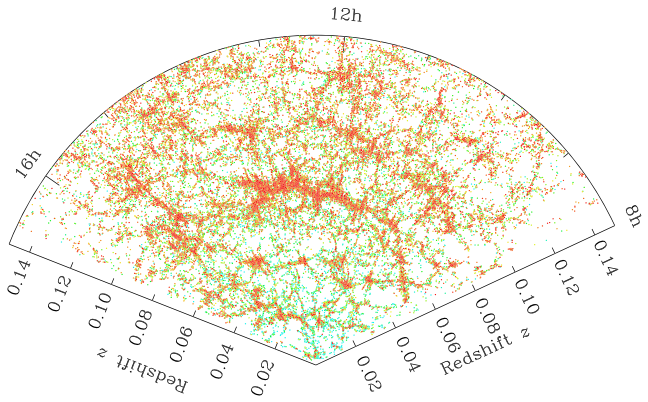



**FIGURE 1.8** A slice through the distribution of the main galaxy sample in the northern part of the SDSS survey, with us the observers situated at the bottom center ($z = 0$). Each dot depicts the position of a galaxy, with color chosen to represent the actual color of the galaxy (i.e., red (dark gray in print version) dots correspond to redder galaxies). Image Credit: Michael Blanton and the Sloan Digital Sky Survey (SDSS) Collaboration.

The galaxies in Fig. 1.8 are clearly not distributed randomly: the universe has structure on large scales. To understand this structure, we must develop the tools to study perturbations around the smooth background. We will see that this is straightforward in theory, as long as the perturbations remain small. To compare theory with observations, we must thus try to avoid regimes that cannot be described by small perturbations. As an extreme example, we can never hope to understand cosmology by carefully examining rock formations on Earth. The intermediate steps—collapse of matter into a galaxy; star formation; planet formation; geology; etc.—are much too complicated to allow comparison between linear theory and observations. In fact, perturbations to the matter on small scales (less than about 10 Mpc) have become large in the late universe; that is, the fractional density fluctuations on these scales are not small, but comparable to or larger than unity. We say that these scales have *grown nonlinear*. On the other hand, large-scale perturbations are still small (quasi-linear). So they have been processed much less than the small-scale structure. Similarly, anisotropies in the CMB are small because they originated at early times and the photons that we observe from the CMB do not clump on their way to us. Because of this, the best ways to learn about the evolution of structure and to compare theory with observations are to look at anisotropies in the CMB and at *large-scale structure* (LSS), i.e. how galaxies and matter are distributed on large scales. However, we will learn in Chs. 12–13 that valuable cosmological information can also be extracted from smaller, nonlinear scales provided we choose our observables wisely.

It is paramount therefore to develop statistics that can empower us to compare maps like that shown in Fig. 1.8 to theories while isolating large scales from small scales. For this purpose, it is often useful to take the Fourier transform of the distribution in question; as we will see, working in Fourier space makes it easier to separate large from small scales. The most important statistic in the cases of both the CMB and the large-scale structure

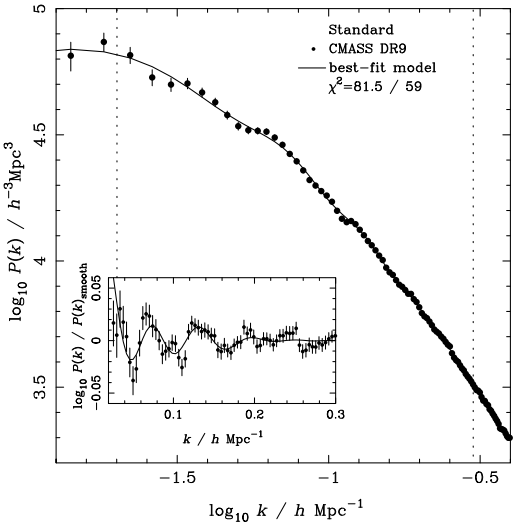



**FIGURE 1.9** The power spectrum $P_{\rm g}(k)$ of the luminous galaxy (CMASS) sample measured in data release 9 of the SDSS-III BOSS survey (points). The solid line is the theoretical prediction from the concordance model (Sect. 1.6), including nonlinear corrections (which we will introduce in Ch. 12). The inset zooms in on the region showing the baryon acoustic oscillation (BAO) feature, which can be used as a standard ruler. From Anderson et al. (2012).

is the *two-point function*, short-hand for two-point correlation function. When measured using Fourier-space fields, it is called the *power spectrum*.

Consider the number density of galaxies in the SDSS survey, as an example. If the density of galaxies as a function of position is $n_{\rm g}(\boldsymbol{x})$, and its mean over the whole survey is $\overline{n}_{\rm g}$, then we can characterize the inhomogeneities with $\delta_{\rm g}(\boldsymbol{x}) = (n_{\rm g}(\boldsymbol{x}) - \overline{n}_{\rm g})/\overline{n}_{\rm g}$, or its Fourier transform $\tilde{\delta}_{\rm g}(\boldsymbol{k})$ (see Box 5.1). By construction, the mean of the field $\delta_{\rm g}(\boldsymbol{x})$ is equal to zero. We then consider the galaxy power spectrum $P_{\rm g}(k)$, which is defined via

$$\langle \tilde{\delta}_{\rm g}(\boldsymbol{k}) \tilde{\delta}^*_{\rm g}(\boldsymbol{k}') \rangle = (2\pi)^3 \delta^{(3)}_{\rm D}(\boldsymbol{k} - \boldsymbol{k}')\, P_{\rm g}(k). \tag{1.10}$$

Here the angular brackets denote an average over the whole ensemble, and $\delta^{(3)}_{\rm D}(\cdot)$ is the Dirac delta function which constrains $\boldsymbol{k} = \boldsymbol{k}'$. The details aside, which we will get to in Ch. 11, Eq. (1.10) indicates that the power spectrum is the spread, or the variance, in the distribution. If there are lots of very under- and overdense regions, the power spectrum will be large, whereas it is small if the distribution is smooth (the power spectrum vanishes identically in a homogeneous universe). Fig. 1.9 shows the measured power spectrum of the galaxy distribution in the SDSS/BOSS survey. We will get to understand its shape, and the interesting oscillatory feature around $k \simeq 0.1\,h$ Mpc$^{-1}$, in Ch. 8.

The best measure of anisotropies in the CMB also is the two-point function, of the intensity on the sky in this case (Ch. 9). There is a technical difference because the CMB temperature is a two-dimensional field, measured everywhere on the sky (i.e., with two angular coordinates but no third coordinate corresponding to distance). Instead of Fourier

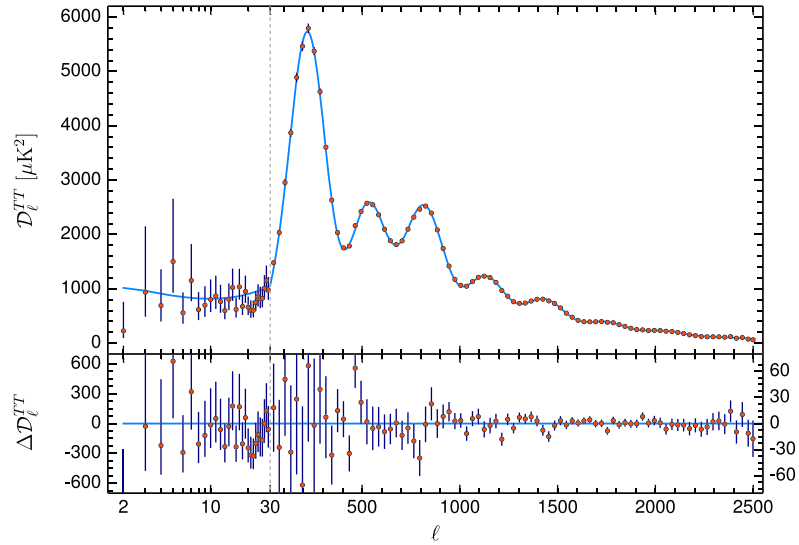



**FIGURE 1.10** *Upper panel:* Anisotropies in the CMB as measured by the Planck satellite (points). The line shows the best-fit prediction by the concordance model of cosmology, based on initial conditions as predicted by inflation. The model involves only six free parameters; its beautiful prediction matches the data almost perfectly. The $x$-axis is multipole moment (e.g., $l = 1$ is the dipole, $l = 2$ the quadrupole) where large angular scales correspond to low $l$; the $y$-axis is the variance of the temperature fluctuations as a function of scale ($\mathcal{D}_l \equiv l(l+1)C(l)T_0^2/2\pi$; we will learn what $C(l)$ is in Ch. 9). The characteristic signature of inflation is the series of peaks and troughs, a signature that has been impressively verified by experiment. *Lower panel:* Difference between data and best-fit model. Notice the change in $y$ axis between $l < 30$ and $l \geq 30$ in this panel. From Planck Collaboration (2018b).

transforming the CMB temperature, then, one typically expands it in spherical harmonics, a basis appropriate for a 2D field on the surface of the sphere. Therefore, the power spectrum of the CMB is a function of multipole moment $l$, not wave number $k$. Dozens of groups have made measurements of the CMB power spectrum since the discovery of anisotropies in 1992. COBE's measurements were at the very largest angles, i.e. low $l$. The definitive measurement was supplied by the Planck satellite in 2018, shown in Fig. 1.10.

One key difference between the map of the CMB and that of the structure in the current universe is the "contrast," or amplitude of structure. The very young universe, as mapped out by CMB experiments, was very smooth, while maps of the current universe as depicted in Fig. 1.8 convince us that the universe is very inhomogeneous today. How did the universe evolve from smooth to clumpy? The simple answer, at the same time one of the most powerful underpinnings of modern cosmology, is that gravity forced more and more matter into overdense regions, so that a region starting out with only a small, $10^{-4}$ fractional overdensity evolved, over billions of years, to become much denser than the homogeneous universe today and in fact the site at which a galaxy formed. During this process, small-scale perturbations grew nonlinear first, and then *hierarchically* assembled to form larger structures.

Fig. 1.9 and Fig. 1.10 both show theoretical curves which are the result of precision calculations within this paradigm of hierarchical gravitational instability, and which agree well with the data. The main goal of much of this book is to develop a first-principle understanding of these theoretical predictions. Indeed, understanding the development of structure in the universe has become the major goal of most cosmologists today. The growth of structure is fundamentally governed by an ongoing competition between gravitational instability, i.e. the tendency of an overdense region to collapse under its own gravity, and the outward pull exerted by the expanding background. Thus, structure is sensitive to the same physics as the background universe itself: its composition, evolution, and curvature. We can then look for consistency between the evolution of the background universe and the growth of structure. This provides a stringent test of our cosmological model as well as the theory that underlies it, general relativity.

While trying to understand the evolution of structure in the universe, we will be forced to confront the question of what generated the initial conditions, that is, the primordial perturbations that were the seeds for this structure. This will lead us to the third important aspect of cosmology (after dark matter and dark energy) that goes beyond the Standard Model of particle physics: the theory of inflation. Chapter 7 introduces this fascinating proposal that the universe expanded exponentially fast when it was but $10^{-35}$ s old. The discoveries of the past two decades have elevated inflation from a theoretical idea with aesthetic appeal to a testable hypothesis. CMB measurements have confirmed most of the basic predictions of inflation, including the absence of spatial curvature.

## 1.1 Standard Model of particle physics

The Standard Model of particle physics describes the known fundamental particles in nature and how they interact. The particles can be divided into two classes: spin-1/2 fermions and integer-spin bosons.

**Fermions** are the constituents of matter: the quarks, out of which baryons are built, and the leptons such as electrons and neutrinos. There are three *generations* with two quarks each for a total of six quarks, denoted $u$, $d$; $s$, $c$; $b$, $t$. Each generation of quarks is associated with a pair of leptons. For example, the $u$, $d$ pair is associated with the electron and its neutrino: $e^-$, $\nu_e$. The other lepton pairs are $\mu^-$, $\nu_\mu$ and $\tau^-$, $\nu_\tau$. The vast majority of matter in the universe is made up of the first generation, with the exception of neutrinos, which are mixed between the different generations. Unlike leptons, quarks do not exist on their own, but they form bound states under the strong interaction. Baryons, the most important ones being the proton and neutron, are made out of three quarks. Mesons are composed of a quark–antiquark pair.

**Bosons** contain the spin-1 (vector) force carriers, the most famous of which is the photon which mediates the electromagnetic force. There are eight gluons (massless, like the photon) that mediate the strong force. The weak force, responsible for example for neutron decay, is mediated by three massive bosons: the $Z$, $W^+$ and $W^-$. These force mediators are complemented with the spin-0 (scalar) Higgs boson. The Higgs couples to all massive fermions as well as the $W$ and $Z$ bosons. This coupling gives mass to the particles through the Higgs' homogeneous background field value.

The Standard Model has remained largely intact since its inception, gaining more and more experimental verification every year. However, neutrino masses are now a confirmed piece of physics beyond the Standard Model. Moreover, the evidence cosmologists have uncovered—

that there is a need for dark matter, dark energy, and new physics leading to inflation—clearly shows that the Standard Model is not the final word in particle physics.

## 1.6 ΛCDM: the concordance model of cosmology

We are now ready to summarize the concordance model of cosmology: a Euclidean universe that is dominated today by non-baryonic cold dark matter (CDM) and a cosmological constant, with initial perturbations generated by inflation in the very early universe. Since all measurements are currently consistent with dark energy being a cosmological constant $\Lambda$, this concordance model of cosmology has become known as (flat) ΛCDM. It is worth noting that none of these ingredients are part of the Standard Model of particle physics (Box 1.1)! Let us thus briefly discuss the status of these three ingredients.

***CDM:*** The "Cold" part of this moniker comes from requiring the dark matter particles to be able to clump efficiently in the early universe. If they are *hot* instead, i.e., have large velocities, structure will not form at the appropriate levels; among others, this excludes the known neutrinos from being dark matter candidates. We have argued that BBN and the CMB imply the existence of non-baryonic matter. However, observations of structure in the universe independently lead to the conclusion that there must be dark matter. The inhomogeneities expected in a model without dark matter are far too small. In Ch. 8, we will come to understand the reason why a baryon-only universe would be so smooth. Moreover, dark matter is a familiar concept to astronomers; the first suggestion was put forth by Zwicky (1933), based on galaxy velocities within clusters. Ample evidence also comes from the rotation curves of galaxies. Indeed, a mismatch between the matter inferred from gravity and that which we can see in the form of baryons exists on all galactic and extragalactic scales, and it always points toward roughly 5 times more dark matter than baryons.

What is this new form of matter? And how did it form in the early universe? So far, we know only its overall abundance and the fact that it must be cold. The most popular idea currently is that the dark matter consists of elementary particles produced during early moments of the Big Bang. In Ch. 4, we will explore this possibility in detail, arguing that dark matter may have been produced when the temperature of the universe was of order hundreds of GeV/$k_{\rm B}$. As we will see, the hypothesis that dark matter consists of fundamental relics from the early universe is being rigorously tested experimentally.

***Cosmological constant:*** Evidence from a variety of sources, but most famously from distant supernovae (starting with Riess et al., 1998; Perlmutter et al., 1999) suggests that there must be energy, *dark energy*, besides ordinary matter and radiation. Unlike dark matter, this component does not cluster strongly. We already discussed the possibility that this new form of energy remains constant with time, i.e., acts as a *cosmological constant*, a possibility first introduced (and later abandoned) by Einstein. Cosmologists have explored other forms though, many of which behave quite differently from the cosmological constant. We will see more of this in Sect. 2.4.6.

A typical physicist confronted with the need to introduce a cosmological constant might at first be quite puzzled: since an expanding universe dilutes the density of particles, it seems impossible to find a fundamental particle that can be the source of dark energy. On the other hand, the notion that empty space itself carries energy, so that the density remains constant even as the universe expands, is consistent with Heisenberg's uncertainty principle and our understanding of quantum mechanics, which has virtual particles and anti-particles popping in and out of existence for brief moments of time, thereby contributing to the vacuum energy. Unfortunately, when one comes to try to quantify the value of the cosmological constant using what we know about quantum field theory and these vacuum fluctuations, the value obtained is much larger than the value required to explain cosmological observations (Exercise 1.5). Dark energy then is more than simply a parameter used to fit the observed universe: it is a supreme puzzle for physics, one that has spawned thousands of papers and ideas, but one that remains unsolved.

***Inflation:*** The most plausible mechanism for generating the initial perturbations that grew into the structure observed today is dubbed inflation. It posits the existence of a brief epoch very early in the universe, during which the scale factor grew exponentially rapidly with time. The epoch of inflation therefore shares some features with our universe today: the dominant form of energy remained roughly constant as the universe expanded, and the identity of the substance driving this rapid expansion is unknown in both cases. However, the scales are much different: the energy density provided by the substance that drove inflation was likely at least 60 orders of magnitude larger than the dark energy density today. Since the energies associated with inflation were likely so large, they are very difficult to probe experimentally. However, we will see that there is at least one signature of inflation that is within reach of experiments and—if detected—would shed light on physics at unprecedented energy scales.

## 1.7 Summary and outlook

As a way of summarizing the features of an expanding universe that we have outlined above and that we will explore in great detail in the coming chapters, let us construct a time line. We can equivalently characterize any epoch in the universe by the time since the Big Bang; by the value of the scale factor at that time; by the redshift freely traveling photons have experienced from then until today or by the temperature of the cosmic background radiation. For example, today, $t \simeq 13.7$ billion years; $a = 1$; $z = 0$; and $T = 2.73\ \mathrm{K} = 2.35 \times 10^{-4}\ \mathrm{eV}/k_B$. Fig. 1.11 shows a time line of the universe using both time and temperature as markers. The milestones indicated on the time line range from those that involve known physics (nucleosynthesis and the CMB) to those that go beyond the Standard Model of particle physics (inflation and dark energy).

The time line in Fig. 1.11 shows the dominant component of the universe at various times. We do not know what dominated the energy budget of the universe at very early times after the end of inflation. We do know, however, that the universe was dominated by radiation at the latest by the time BBN occurred. Eventually, since the energy of a relativistic particle falls as $1/a$ while that of a nonrelativistic particle remains constant at $m$, matter

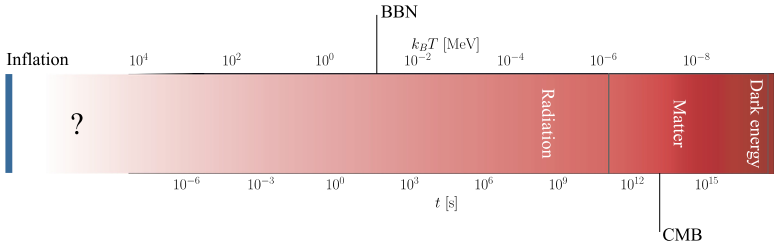



**FIGURE 1.11** A history of the universe. Any epoch can be associated with either temperature (top scale) or time (bottom scale). Also indicated are the types of constituents that dominate at any given time. At very early times, we do not know whether the universe was always radiation dominated.

overtook radiation. At relatively recent times, the universe has become dominated not by matter, but by dark energy, whose density remains approximately constant with time.

The classical results in cosmology can be understood in the context of a smooth universe. Light elements formed when the universe was several minutes old, and the CMB decoupled from matter at a temperature of order $k_{\rm B}T \sim 1/4$ eV, when the universe was 380,000 years old. Heavy elementary particles may make up the dark matter in the universe; if they do, their abundance was fixed at very high temperatures of order $k_{\rm B}T \sim 100$ GeV or higher.

In this book, we will be mostly interested in the perturbations around the smooth universe. At the beginning of the time line, we allow for a brief period of inflation, during which primordial perturbations were produced. These small perturbations began to grow when the universe became dominated by matter. The dark matter grew more and more clumpy, simply because of the attractive nature of gravity. An overdensity of dark matter of 1 part in 1000 when the temperature was 1 eV grew to 1 part in 100 by the time the temperature dropped to 0.1 eV. Eventually, at relatively recent times, perturbations in the matter ceased to be small; they became the nonlinear structure we see today. The observed anisotropies in the CMB tell us what the universe looked like when perturbations were very small, so they are a wonderful probe of the latter. Moreover, the CMB anisotropies provide a precise characterization of the initial conditions needed for detailed analytic and numerical studies of the growth of structure. To give you an idea of the road ahead, Fig. 1.12 charts the way through the various ingredients going into this calculation that we will get to know in subsequent chapters of the book.

Some of the elements in the time line we have discussed may well be incorrect. However, since most of these ideas are testable, the data from the first half of the 21st century will tell us which parts of the time line are correct and which need to be discarded. This in itself seems more than sufficient reason to study the CMB and large-scale structure.

## Exercises

**1.1** Suppose (incorrectly) that $H$ scales as temperature squared all the way back until the time when the temperature of the universe was $10^{19}$ GeV/$k_{\rm B}$ (i.e., suppose the uni-

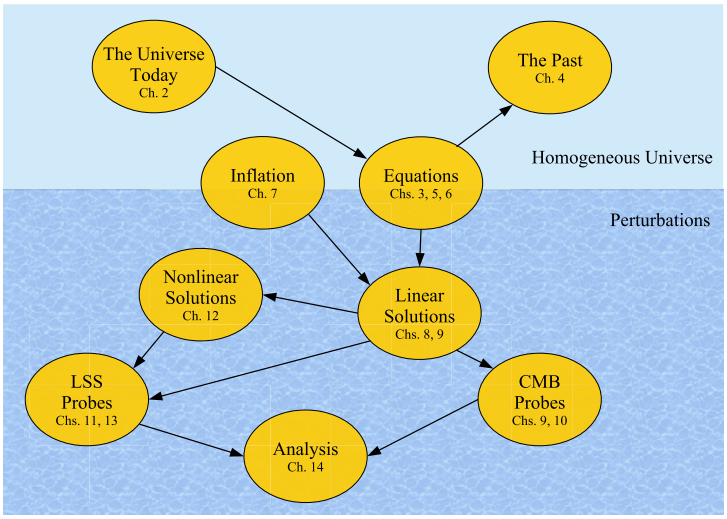



**FIGURE 1.12** An outline of the remainder of this book. The homogeneous universe is mostly described in Chs. 2–4, while Chs. 5–6 derive the equations governing the (linear) inhomogeneous universe. Ch. 7, which starts with some mysteries about the homogeneous universe and motivates the theory of inflation, also describes the generation of perturbations, which we then follow forward by solving the perturbation equations (Chs. 8–9). This allows us to predict the CMB (Chs. 9–10) and obtain results on large-scale structure (LSS), such as galaxy clustering (Ch. 11). We then study the *nonlinear* evolution of perturbations to the matter in Ch. 12, before turning to gravitational lensing in Ch. 13. The final chapter of the book is devoted to how we go about extracting cosmological information from data based on the predictions we derived throughout the book. The arrows indicate how each chapter builds on results of previous chapters.

verse was radiation dominated all the way back to the *Planck time*). Also suppose that today the dark energy is in the form of a cosmological constant $\Lambda$, such that $\rho_\Lambda$ today is equal to $0.7\rho_{\rm cr}$ and $\rho_\Lambda$ remains constant throughout the history of the universe. What was $\rho_\Lambda/(3H^2/8\pi G)$ back then? The purpose of this estimate is to illustrate how odd (or unnatural) is the value of $\Lambda$. For a variety of reasons (see Exercise 1.5), one would expect its natural value to yield an energy density comparable to the ambient density at the Planck time.

**1.2** Assume the universe today is Euclidean with both matter and a cosmological constant, the latter with energy density that remains constant with time. Integrate Eq. (1.2) to find the present age of the universe (since radiation dominated only during a small fraction of the universe's age, including only matter and the cosmological constant is a good approximation). That is, assume that $\rho(t_0) = \rho_{\rm cr}$ and use Eq. (1.3) to write

$$dt = H_0^{-1}\frac{da}{a}\left[\Omega_\Lambda + \frac{1-\Omega_\Lambda}{a^3}\right]^{-1/2} \tag{1.11}$$

where $\Omega_\Lambda$ is the ratio of energy density in the cosmological constant to the critical density (see Eq. (2.71)). Integrate from $a = 0$ (when $t = 0$) until today at $a = 1$ to get the age of the universe today. In both cases below the integral can be done analytically.

**(a)** First do the integral in the case when $\Omega_\Lambda = 0$.

**(b)** Now do the integral in the case when $\Omega_\Lambda > 0$. *Hint:* Define a new integration variable $x \equiv \ln(1/a^3)$ and then use the fact that

$$\int \frac{dx}{\sqrt{1+\alpha e^x}} = -2\coth^{-1}\left(\sqrt{\alpha e^x + 1}\right). \tag{1.12}$$

**(c)** For fixed $H_0$, which universe is older?

**1.3** Using the reduced masses of hydrogen and deuterium, and the fact that the Lyman-$\alpha$ ($n = 2 \to n = 1$) transition in H has a wavelength 1216 $\mathring{A}$, find the wavelength of the photon emitted in the corresponding transition in D. Astronomers often define $c\Delta\lambda/\lambda$ to characterize the splitting of two nearby lines. What is this quantity for the H–D pair?

**1.4** Convert the specific intensity in Eq. (1.9) into an expression for what is plotted in Fig. 1.7, the energy per area, time, frequency and steradian. Show that the peak of a 2.73 K black-body spectrum does lie at $1/\lambda \simeq 5\ \mathrm{cm}^{-1}$. What frequency does this correspond to?

**1.5** The ground state energy of the harmonic oscillator is $\hbar\omega/2$. This ground state energy carries through to quantum field theory where the fluctuations of fields even in empty space lead to an energy density equal to

$$\rho_{\rm vacuum} = \int \frac{d^3p}{(2\pi\hbar)^3}\frac{\hbar\omega}{2} \tag{1.13}$$

where the integral corresponds to a sum over all possible momentum modes and, for a particle with mass $m$, the energy is equal to $\hbar\omega = \sqrt{m^2c^4 + p^2c^2}$. The integral diverges, but this simply reflects the fact that above a certain scale $E_{\rm max} = p_{\rm max}c$, there is likely to be new physics that changes the fundamental degrees of freedom. Calculate this integral for electrons for two values: $E_{\rm max} = 10m_ec^2$ (which is extremely conservative, because we certainly know about physics above this energy scale) and $E_{\rm max} = m_{\rm Pl}c^2 = 1.2 \times 10^{19}$ GeV (the scale above which quantum mechanical corrections to gravity become large). Compare the values obtained with the value of the dark energy density today $\rho_\Lambda \simeq 3 \times 10^{-11}\ \mathrm{eV}^4/(\hbar c)^3$. *Caveat:* There is some disagreement in the literature as to whether this way of handling the divergence is correct (Martin, 2012); however, everyone agrees that there is a big problem.

# 2

# The expanding universe

Just as the early navigators of the great oceans required sophisticated tools to help them find their way, we will need modern technology to help us work through the ramifications of an expanding universe. In this chapter, we introduce the metric and the distribution function, the first of which underlies general relativity and the second, statistical mechanics. We will use this language to derive some of the basic features of the smooth, expanding universe: the redshifting of light, the notion of distance needed to understand the arguments for dark energy, the evolution of the energy density with scale factor, and the epoch of equality $a_{\rm eq}$ shown in Fig. 1.3. We then go on to perform a cosmic inventory, identifying those constituents of the universe that dominate the energy budget at various epochs.

Implicit in this discussion will be the notion that the universe is smooth, more precisely: spatially homogeneous. That is, the densities of the various constituents such as matter and radiation do not vary in space. To make things even simpler, we will work under the assumption—which is observed to be correct and the reason for which is understood—that all the constituents have equilibrium distributions, as defined and explored in Sect. 2.3.

These simple assumptions form the basic framework within which cosmologists operate and around which they perturb, so that a good grasp of this "zeroth-order universe" is essential. In subsequent chapters, we will see that the deviations from smoothness and the equilibrium distributions are the source of much of the richness we observe in the universe.

From this chapter onward, we use units in which

$$\hbar = c = k_{\rm B} = 1. \tag{2.1}$$

Many research papers employ these units, so it is important to get accustomed to them. Please work through Exercise 2.1 if you are uncomfortable with the idea of setting the speed of light, or Planck's and Boltzmann's constants to 1.

## 2.1 Expanding space

On the one hand, the expansion of the universe and the perturbations that make it interesting are governed by general relativity. On the other hand, most of cosmology can be learned with only a passing knowledge of this formidable theory. Very roughly, we can break down relativity into two parts. The first is the idea of general covariance: any physical phenomenon can be described within any desired frame of reference, with the same result. Einstein realized that this principle requires that space and time (unified through special

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00008-5
Copyright © 2021 Elsevier Inc. All rights reserved.

relativity) are *curved*. To describe a curved spacetime, we introduce the metric, which tells us the distance between two points in time and space. The second part of general relativity relates spacetime, via the metric, to all the stuff that is contained within it, such as matter and radiation. This part of relativity is summarized in the form of Einstein's equations, which we will introduce in Ch. 3. In this chapter, we will deal only with expanding space and the metric that describes it, which are aspects of the first part of Einstein's theory, and independent of Einstein's equations.

### 2.1.1 The metric

Rigorously defined, the *metric* returns the actual physical distance between two infinitesimally close points in spacetime defined in some arbitrary coordinate system. It will be an essential tool in our quest to make quantitative predictions in an expanding universe. In fact, long before Einstein, physicists such as Newton and Maxwell used a spacetime metric. However, their use of a metric was implicit, since they did not distinguish between space and the coordinates that describe it. Going back to Fig. 1.1 from Ch. 1, we see that even if one knows the components of a separation vector between two points, say two grid points in that figure, the physical distance associated with this vector requires additional information; in this case, the value of the scale factor $a(t)$ at that time.

We are familiar with the metric for the Cartesian coordinate system $(x, y)$ which says that the square of the physical distance between two points separated by $dx$ and $dy$ in a 2D plane is $(dx)^2 + (dy)^2$. However, if we use polar coordinates $(r, \theta)$ instead, the square of the physical distance no longer is the sum of the square of the two coordinate differences. Rather, if the differences $dr$ and $d\theta$ are small, the square of the distance between two points is $(dr)^2 + r^2(d\theta)^2$. This distance is *invariant*: an observer using Cartesian coordinates to calculate it would get the same result as one using polar coordinates. Thus another way of stating what a metric does is this: it turns observer-dependent coordinates into invariants. Mathematically, in the 2D plane, the invariant distance squared is $dl^2 = \sum_{i,j=1,2} g_{ij} dx^i dx^j$. The metric $g_{ij}$ in this 2D example is a $2 \times 2$ symmetric matrix. In Cartesian coordinates $(x_1 = x, x_2 = y)$ the metric is simply the identity matrix

$$g_{ij} \overset{\text{Cartesian}}{=} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.2}$$

while in polar coordinates $(x_1 = r, x_2 = \theta)$ it instead becomes

$$g_{ij} \overset{\text{polar}}{=} \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}. \tag{2.3}$$

Note that $g_{ij}$ can also depend on location (in this case through $r$). Both forms of the metric describe the same space: a 2D plane.

The concept of a metric really comes into its own when considering more general, curved spaces. Consider the surface of the Earth, which we can roughly approximate as a sphere. There are various ways to assign coordinates to a point on the Earth's surface.

Plotting them on a flat piece of paper results in a *map*. But is there a coordinate system such that the resulting map accurately reflects distances, areas, and angles, like the Cartesian coordinates do for Euclidean space? The answer is no: the Earth's surface is curved, and no such coordinate representation exists. Fig. 2.1 illustrates this. The Mercator coordinates (top panel) preserve angles,[1] which make them useful for navigation. But they strongly distort *apparent* distances and areas, especially near the poles. Another choice are the Winkel–Tripel coordinates, which reduce the apparent distortion of distances and areas, but instead feature a distortion of angles. This apparent problem for map-making does not pose a real problem for us, however: we know that we must use the metric to calculate distances, areas, and angles. And while the metric looks different in different coordinates, the result will be independent of which coordinates we use.

For us as physicists, another reason we use the metric is that, by describing curved spacetime, it incorporates gravity. Instead of thinking of gravity as an external force and talking of particles moving in a gravitational field, we can include gravity in the metric and talk of particles moving freely in a distorted or curved spacetime. The underlying principle is again general covariance: as Einstein realized, an observer in a uniform gravitational field makes exactly the same measurements as one in an accelerated reference frame.

In four spacetime dimensions the invariant includes time intervals as well, so that

$$ds^2 = \sum_{\mu,\nu=0}^{3} g_{\mu\nu} dx^\mu dx^\nu \tag{2.4}$$

where the indices $\mu$ and $\nu$ range from 0 to 3 (see Box 2.1), with the first one reserved for the time-like coordinate (e.g., $x^0 = t$) and the last three for spatial coordinates. As in special relativity, the time-time component of the metric has the opposite sign of the purely spatial components. Here, we will choose the "mostly positive" metric convention, where the spatial metric components are positive, following standard convention in cosmology. In Eq. (2.4) we have explicitly written down the summation sign, but from now on we will use the convention that repeated indices are summed over. $g_{\mu\nu}$ is symmetric, so it has four diagonal and six independent off-diagonal components.

The metric provides the connection between values of the coordinates and the physical measure of the interval $ds^2$. This interval is often called the *proper-time* interval. To see why, imagine an observer sitting with a watch. This observer naturally chooses a coordinate system $\{t, \boldsymbol{x}\}$ within which she remains at the origin and where the time coordinate corresponds to the time shown by her watch. Now define two spacetime events as the points where the observer's watch shows 12:00:00 and 12:00:01. Since the observer does not move with respect to the $x^i$, we have $dx^i = 0$ and thus the invariant interval becomes $ds^2 = g_{00} dt^2 = -(1\,\mathrm{s})^2$. Thus, apart from the minus sign, the proper-time interval is precisely the time elapsed according to the observer's watch. Another observer, moving with

[1] This means that the angle between straight lines connecting a given point on the map with two other nearby points is the same as would be measured on the actual surface of the Earth.

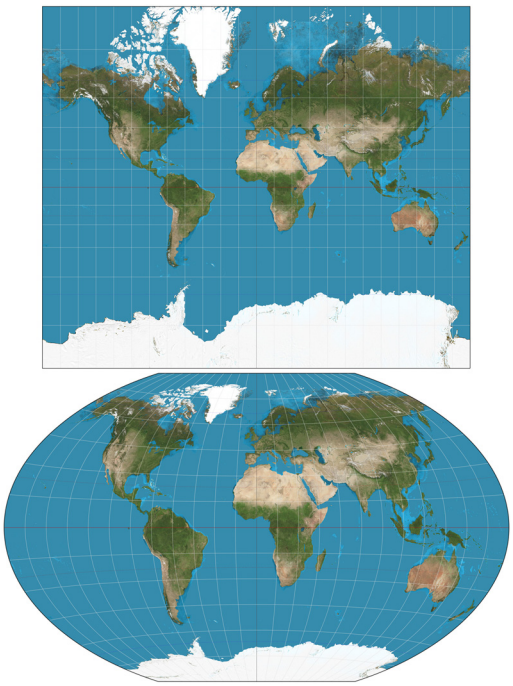


**FIGURE 2.1** The surface of the Earth represented in two different coordinate systems: Mercator coordinates (top) and Winkel–Tripel coordinates (bottom). The Mercator coordinates preserve angles but visually greatly distort areas: Greenland appears to be larger in area than Australia, despite having less than a third of Australia's area in reality. The Winkel–Tripel coordinates improve this, at the price of distorting angles. No flat representation of the Earth's surface can be faithful in both areas and angles, because it is curved. The metric, however, allows us to correctly compute distances, areas, and angles regardless of the coordinates chosen. Image credit: Daniel R. Strebe (2011), CC BY-SA 3.0.

respect to the first, may assign a different $dt$ and $dx^i$ but will compute an identical value of the proper-time interval $ds$ for the same two events. The negative sign means that the pair of events are separated by a time-like interval, while events that have a positive proper-time interval are separated by a space-like interval; two events with $ds^2 = 0$ are connected by light rays.

## 2.1 Indices

In three dimensions, a vector $\boldsymbol{A}$ has three components, which we refer to as $A^i$, with the superscript $i$ taking the values 1, 2, or 3. The dot product of two vectors is then

$$\boldsymbol{A} \cdot \boldsymbol{B} = \sum_{i=1}^{3} A^i B^i \equiv A^i B^i \tag{2.5}$$

where we have introduced the Einstein summation convention of not explicitly writing the $\sum$ sign when an index (in this case $i$) appears twice. Similarly, matrices can be written in compo-

nent notation. For example, the product of two matrices $M$ and $N$ is

$$(MN)_{ij} = M_{ik}N_{kj} \tag{2.6}$$

again with implicit summation over $k$.

In relativity, two generalizations must be made. First, every vector has a fourth component, the time component. Since the spatial indices run from 1 to 3, it is conventional to use 0 for the time component. Greek letters are used to represent all four components, so $A^\mu = (A^0, A^i)$. The second, more subtle, feature of relativity is the distinction between upper and lower indices, the former associated with *contravariant*, the latter with *covariant* vectors. One goes back and forth with the metric tensor, so that

$$A_\mu = g_{\mu\nu}A^\nu; \quad A^\mu = g^{\mu\nu}A_\nu. \tag{2.7}$$

A contravariant vector and a covariant vector can be contracted to produce an invariant, a scalar. For example, the statement that the four-momentum squared of a massless particle must vanish is

$$P^2 \equiv P_\mu P^\mu = g_{\mu\nu}P^\mu P^\nu = 0. \tag{2.8}$$

Just as the metric can turn an upper index on a vector into a lower index, the metric can be used to raise and lower indices on tensors with an arbitrary number of indices. For example, raising the indices on the metric tensor itself leads to

$$g^{\mu\nu} = g^{\mu\alpha}g^{\nu\beta}g_{\alpha\beta}. \tag{2.9}$$

If the index $\alpha = \nu$, then the first term on the right is equal to the term on the left, so if the combination of the last two terms on the right force $\alpha$ to be equal to $\nu$, then the equation is satisfied. Therefore,

$$g^{\nu\beta}g_{\beta\alpha} = \delta^\nu_{\ \alpha}, \tag{2.10}$$

where $\delta^\nu_{\ \alpha}$ is the Kronecker delta, which is equal to the identity matrix: identical to zero unless $\nu = \alpha$ in which case it is 1. Thus, $g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$.

Special relativity is described by the *Minkowski spacetime* with the metric: $g_{\mu\nu} = \eta_{\mu\nu}$, where

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.11}$$

This is the metric implicitly used by Maxwell in deriving his equations of electromagnetism. It describes a spacetime that is not curved.

Now, what is the metric that describes the expanding universe? Let us return to the grid depicted in Fig. 1.1. Two grid points move away from each other, such that the distance between the two points is always proportional to the scale factor $a(t)$. If the coordinate (i.e., comoving) distance today is $x_0$, the physical distance between the two points at some earlier time $t$ was $a(t)x_0$ with $a$ today, $a_0$, equal to one. At least in a Euclidean (or "flat,"

as opposed to open or closed) universe, the metric then is *almost* the Minkowski metric, except that spatial coordinates must be multiplied by the scale factor. This suggests that the metric in an expanding, Euclidean universe is

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & a^2(t) & 0 & 0 \\ 0 & 0 & a^2(t) & 0 \\ 0 & 0 & 0 & a^2(t) \end{pmatrix}. \tag{2.12}$$

This is the Friedmann–Lemaître–Robertson–Walker (FLRW) metric for a Euclidean universe.

In order to determine how the function $a(t)$ evolves with time, we need to know the composition of the homogeneous constituents in the universe, and we need to use Einstein's equations, as noted in Eq. (1.3). We will turn to that in Sect. 3.1. When perturbations are introduced, the metric will become more complicated and will in addition depend on the location in space. Therefore, Eq. (2.12) will be generalized to include functions that depend on both time and space and quantify deviations from uniformity. These perturbed parts of the metric will be determined by the inhomogeneities in the matter and radiation.

Before that, however, let us consider how matter and radiation behave within an expanding spacetime, and how we can go from infinitesimal invariant intervals to actual finite distances.

### 2.1.2 The geodesic equation

In Minkowski space, particles travel in straight lines unless they are acted on by a force. Not surprisingly, the paths of particles in more general spacetimes are more complicated. In a curved space, the notion of a straight line gets generalized to a *geodesic*, the shortest path (or, in general extremal path) between two points. Quite beautifully, general relativity states that this is precisely the path followed by a particle in the absence of any forces apart from gravity. To express this in equations, we must generalize Newton's law with no forces, $d^2\boldsymbol{x}/dt^2 = 0$, to accommodate more general coordinate systems and spacetimes.

The machinery necessary to generalize $d^2\boldsymbol{x}/dt^2 = 0$ is perhaps best introduced by starting with a simple example: free particle motion in a Euclidean 2D plane. In that case, the equations of motion in Cartesian coordinates $x^i = (x, y)$ are

$$\frac{d^2x^i}{dt^2} = 0. \tag{2.13}$$

However, if we use polar coordinates $x'^i = (r, \theta)$ instead, the equations for a free particle look significantly different. The fundamental difference between the two coordinate systems is that the basis vectors for polar coordinates $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{\theta}}$ vary in the plane. Therefore, the coordinates $r$ and $\theta$ *do not* satisfy $d^2x'^i/dt^2 = 0$.

To determine the equation satisfied by a free particle in polar coordinates, we can start from the Cartesian equation and then transform. In particular,

$$\frac{dx^i}{dt} = \frac{\partial x^i}{\partial x'^j}\frac{dx'^j}{dt}. \tag{2.14}$$

$\partial x^i/\partial x'^j$ is called the *transformation matrix* going from one coordinate system to another. In the case of Cartesian to polar coordinates in 2D, $x^1 = x'^1\cos x'^2$ and $x^2 = x'^1\sin x'^2$, so the transformation matrix is

$$\frac{\partial x^i}{\partial x'^j} = \begin{pmatrix} \cos x'^2 & -x'^1\sin x'^2 \\ \sin x'^2 & x'^1\cos x'^2 \end{pmatrix}. \tag{2.15}$$

Therefore, the geodesic equation becomes

$$\frac{d}{dt}\left[\frac{dx^i}{dt}\right] = \frac{d}{dt}\left[\frac{\partial x^i}{\partial x'^j}\frac{dx'^j}{dt}\right] = 0. \tag{2.16}$$

The derivative with respect to time acts on both terms inside the brackets. If the derivative acting on the transformation matrix vanished, the geodesic equation in the new coordinates would still be $d^2x'^i/dt^2 = 0$. In the case of polar coordinates, though, this derivative does not vanish, and we can apply the chain rule to obtain

$$\frac{d}{dt}\left(\frac{\partial x^i}{\partial x'^j}\right) = \frac{\partial^2 x^i}{\partial x'^j\partial x'^k}\frac{dx'^k}{dt}. \tag{2.17}$$

The geodesic equation in the new coordinates therefore becomes

$$\frac{d}{dt}\left[\frac{\partial x^i}{\partial x'^j}\frac{dx'^j}{dt}\right] = \frac{\partial x^i}{\partial x'^j}\frac{d^2x'^j}{dt^2} + \frac{\partial^2 x^i}{\partial x'^j\partial x'^k}\frac{dx'^k}{dt}\frac{dx'^j}{dt} = 0. \tag{2.18}$$

To get this in a more recognizable form, note that the term multiplying the second time derivative $d^2x'^j/dt^2$ is the transformation matrix. If we multiply the equation by the inverse of this transformation matrix, then the second time derivative will stand alone, leaving

$$\frac{d^2x'^l}{dt^2} + \left[\left(\left\{\frac{\partial x}{\partial x'}\right\}^{-1}\right)^l_i \frac{\partial^2 x^i}{\partial x'^j\partial x'^k}\right]\frac{dx'^k}{dt}\frac{dx'^j}{dt} = 0. \tag{2.19}$$

You can check that this rather cumbersome expression does indeed give the correct equations of motion in polar coordinates. This is the geodesic equation in a non-Cartesian coordinate system.

It is convenient to define the *Christoffel symbol*, $\Gamma^l{}_{jk}$, to be the coefficient in brackets in Eq. (2.19). Note that by definition it is symmetric in its lower indices $j$ and $k$. In a Cartesian coordinate system describing a Euclidean space, the Christoffel symbol vanishes and the geodesic equation is simply $d^2x^i/dt^2 = 0$. But in general, the Christoffel symbol does not vanish; its presence describes geodesics in nontrivial coordinate systems. In a nontrivial

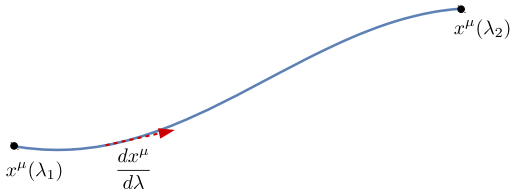



**FIGURE 2.2** A particle's path $x^\mu(\lambda)$ is parametrized by $\lambda$, which monotonically increases from its initial value $\lambda_1$ to its final value $\lambda_2$. The tangent to the path is given by the vector $dx^\mu/d\lambda$ (arrow).

spacetime such as the expanding universe it is *not possible* to find a coordinate system in which the Christoffel symbol vanishes identically, so the geodesic equation is essential.

There are two small changes we need to make when importing the geodesic equation (2.19) into relativity. The first is trivial: allow the indices to range from 0 to 3 to include time and the three spatial dimensions. The second is also not too surprising: since time is now one of our coordinates, fixing the evolution parameter to be the time coordinate does not always work (although it certainly is a possible choice in many cases). Instead, we introduce a parameter $\lambda$ which monotonically increases along the particle's path as in Fig. 2.2. The geodesic equation then becomes

$$\frac{d^2x^\mu}{d\lambda^2} + \Gamma^\mu{}_{\alpha\beta}\frac{dx^\alpha}{d\lambda}\frac{dx^\beta}{d\lambda} = 0. \tag{2.20}$$

We derived this equation transforming from a Cartesian basis, so that the Christoffel symbol is given by the term in square brackets in Eq. (2.19). It is almost always more convenient, however, to obtain the Christoffel symbol from the metric directly. A convenient formula expressing this dependence is

$$\Gamma^\mu{}_{\alpha\beta} = \frac{g^{\mu\nu}}{2}\left[\frac{\partial g_{\alpha\nu}}{\partial x^\beta} + \frac{\partial g_{\beta\nu}}{\partial x^\alpha} - \frac{\partial g_{\alpha\beta}}{\partial x^\nu}\right]. \tag{2.21}$$

Note again that the raised indices on $g^{\mu\nu}$ are important: $g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$ (see Box 2.1). So $g^{\mu\nu}$ in the Euclidean FLRW metric is similar to $g_{\mu\nu}$, the only difference being that its spatial elements are $1/a^2$ instead of $a^2$.

To understand how particles move in an expanding universe, then, we first need to calculate the Christoffel symbol. The starting points are the general expression in Eq. (2.21) and the FLRW metric in Eq. (2.12). First we compute the components with upper index equal to zero, $\Gamma^0{}_{\alpha\beta}$. Since the metric is diagonal, the factor of $g^{0\nu}$ vanishes unless $\nu = 0$ in which case it is $-1$. Therefore,

$$\Gamma^0{}_{\alpha\beta} = -\frac{1}{2}\left[\frac{\partial g_{\alpha 0}}{\partial x^\beta} + \frac{\partial g_{\beta 0}}{\partial x^\alpha} - \frac{\partial g_{\alpha\beta}}{\partial x^0}\right]. \tag{2.22}$$

The first two terms here reduce to derivatives of $g_{00}$. Since the FLRW metric has constant $g_{00}$, these terms vanish, and we are left with

$$\Gamma^0{}_{\alpha\beta} = \frac{1}{2}\frac{\partial g_{\alpha\beta}}{\partial x^0}. \tag{2.23}$$

The derivative is nonzero only if $\alpha$ and $\beta$ are spatial indices, which will be identified with Roman letters $i$, $j$ running from 1 to 3. Since $x^0 = t$, we have

$$\begin{aligned}
&\Gamma^0{}_{00} = 0,\\
&\Gamma^0{}_{0i} = \Gamma^0{}_{i0} = 0,\\
&\Gamma^0{}_{ij} = \delta_{ij}\,\dot{a}\,a.
\end{aligned} \tag{2.24}$$

It is a straightforward and useful exercise to show that $\Gamma^i{}_{\alpha\beta}$ is nonzero only when one of its lower indices is zero and one is spatial, so that

$$\Gamma^i{}_{0j} = \Gamma^i{}_{j0} = \delta_{ij}\frac{\dot{a}}{a} \tag{2.25}$$

with all other $\Gamma^i{}_{\alpha\beta}$ zero.

This has been a long, rather formal subsection, opening with the generalization of the geodesic equation to curved spacetime and then proceeding with a calculation of the Christoffel symbol in the expanding universe described by the FLRW metric. We can now enjoy the fruits of our labor by applying this formalism to a single particle. In particular, let us see how a particle's energy changes as the universe expands. We will do the calculation here for a massless particle; an almost identical problem for a massive particle is relegated to Exercise 2.3.

Start with the four-dimensional energy-momentum vector $P^\alpha = (E, \boldsymbol{P})$, whose time component is the energy. We use this four-vector to define the parameter $\lambda$ in Eq. (2.20):

$$P^\alpha = \frac{dx^\alpha}{d\lambda}. \tag{2.26}$$

This is an implicit definition of $\lambda$. Fortunately, one never needs to find $\lambda$ explicitly, for it can be directly eliminated by noting that

$$\begin{aligned}
\frac{d}{d\lambda} &= \frac{dx^0}{d\lambda}\frac{d}{dx^0}\\
&= E\frac{d}{dt}.
\end{aligned} \tag{2.27}$$

The reason we define $\lambda$ as the affine parameter in this way is that it allows us to treat both massive and massless particles. For massive particles, the proper time (i.e. what the watch of an observer traveling along with the particle would show) is a more intuitive choice, but this does not work for massless particles, since they travel along geodesics with vanishing

proper-time interval, $ds = 0$. In any case, the trajectories of particles are independent of what affine parameter we choose to describe them.

In the FLRW metric, the 0-component of the geodesic equation (2.20) then becomes

$$E\frac{dE}{dt} = -\Gamma^0{}_{ij}P^iP^j \tag{2.28}$$

where the equality holds since only the spatial components of $\Gamma^0{}_{\alpha\beta}$ are nonzero. Inserting these components leads to a right-hand side equal to $-\delta_{ij}a\dot{a}P^iP^j$. A massless particle has energy-momentum vector $(E, \boldsymbol{P})$ with zero magnitude:

$$g_{\mu\nu}P^\mu P^\nu = -E^2 + \delta_{ij}a^2P^iP^j = 0, \tag{2.29}$$

which enables us to write the right-hand side of Eq. (2.28) as $-(\dot{a}/a)E^2$. Therefore, the geodesic equation yields

$$\frac{dE}{dt} + \frac{\dot{a}}{a}E = 0, \tag{2.30}$$

the solution to which is

$$E = \frac{E_0}{a}. \tag{2.31}$$

This confirms our hand-waving argument in Ch. 1 that the energy of a massless particle should decrease as the universe expands since it is inversely proportional to its wavelength, which is being stretched along with the expansion. In Ch. 3 we will rederive this result in yet another way using the Boltzmann equation.

One final comment about the relation between energy and momentum for a massless particle as expressed in Eq. (2.29). If we define

$$p^i = aP^i, \tag{2.32}$$

we have $E^2 = \delta_{ij}p^ip^j$, so that we can identify $\boldsymbol{p}$ with the physical momentum (while $P^i$ is the momentum defined with respect to the comoving grid). In terms of the physical momentum, the well-known relation

$$E = p \quad \text{where} \quad p \equiv |\boldsymbol{p}| \tag{2.33}$$

continues to hold, which is why we will mostly use this version of particle momentum.

## 2.2 Distances

We can anticipate that measuring distance in an expanding universe will be a tricky business. Referring back to the expanding grid of Fig. 1.1, we immediately see two possible ways to measure distance, the comoving distance which remains fixed as the universe expands or the physical distance which grows simply because of the expansion. Frequently,

neither of these two measures accurately describes the process of interest. For example, light leaving a distant galaxy at redshift 3 starts its journey towards us when the scale factor was only a quarter of its present value and ends it today when the universe has expanded by a factor of 4. Which distance do we use in that case to relate, say, the luminosity of the galaxy to the flux we see?

The starting point for the calculation of distances is the *comoving* (or coordinate) distance which refers to the coordinate grid and is simple to define mathematically. Consider the comoving distance between a distant light source and us. In a small time interval $dt$, light travels a comoving distance $dx = dt/a$ (recall that we are setting $c = 1$), so the total comoving distance traveled by light that began its journey from an object at time $t$ when the scale factor was equal to $a$ (or redshift $z = 1/a - 1$) is

$$\chi(t) = \int_t^{t_0} \frac{dt'}{a(t')} = \int_{a(t)}^{1} \frac{da'}{a'^2 H(a')} = \int_0^z \frac{dz'}{H(z')}. \tag{2.34}$$

Here we have changed the integration over $t'$ to one over $a'$, which brings in the additional factor of $\dot{a} = aH$ in the denominator, and finally to $z'$. As the final expression makes clear, for small redshifts $z$ we can write the comoving distance as $\chi \approx z/H_0$ (verifying our hand-waving discussion of the Hubble diagram at small redshifts in Sect. 1.2). The behavior at larger redshift in the fiducial concordance cosmology is depicted in Fig. 2.3.

Before relating the comoving distance to observables, let us take a quick detour to consider the comoving distance $\eta$ that light could have traveled (in the absence of interactions) since $t = 0$,

$$\eta(t) \equiv \int_0^t \frac{dt'}{a(t')}. \tag{2.35}$$

The reason this distance is so important is that no information could have propagated further on the coordinate grid than $\eta$ since the beginning of time. Therefore, regions separated by distances greater than $\eta$ are not causally connected. If they appear similar, we should be suspicious! We can think of $\eta$ then as the *comoving horizon*. We can also think of $\eta$, which is monotonically increasing, as a time variable and call it the *conformal time*. Just like the time $t$, the temperature $T$, the redshift $z$, and the scale factor $a$, we can use $\eta$ as time variable to describe the evolution of the universe. In fact, for most purposes $\eta$ is the most convenient time variable, so when we begin to study the evolution of perturbations, we will use it instead of $t$. In some simple cases, $\eta$ can be expressed analytically in terms of $a$ (Exercise 2.6). For example, in a matter-dominated universe, $\eta \propto a^{1/2}$, while $\eta \propto a$ in a radiation-dominated universe.

A classic way to determine distances in astronomy is to measure the angle $\theta$ subtended by an object of known physical size $l$ ("standard ruler"). Since this angle is small (almost always in astronomy), the distance to that object is

$$d_A = \frac{l}{\theta}. \tag{2.36}$$

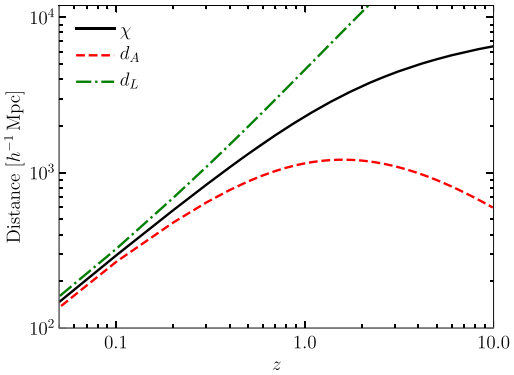



**FIGURE 2.3** Three distance measures in the Euclidean expanding universe in the concordance cosmological model: the comoving distance $\chi$, the angular diameter distance $d_A$, and the luminosity distance $d_L$.

This relation defines the *angular diameter distance* $d_A$. To compute the angular diameter distance in an expanding universe, we first note that the comoving size of the object is $l/a$, where $a$ is the scale factor when the light is emitted. The comoving distance out to the object is given by Eq. (2.34), so the angle subtended in a Euclidean universe is $\theta = (l/a)/\chi(a)$. Comparing with Eq. (2.36), we see that the angular diameter distance is

$$d_A^{\rm Euc} = a\,\chi = \frac{\chi}{1+z}. \tag{2.37}$$

Note that the angular diameter distance is equal to the comoving distance at low redshift, but actually decreases at very large redshift (Fig. 2.3). At least in a Euclidean universe, objects at large redshift appear larger than they would at intermediate redshift! This is a consequence of the fact that the entire universe was smaller (in the sense that each grid cell in Fig. 1.1 was smaller), and hence the emitting galaxy and we observers were physically much closer.

The superscript "Euc" is a warning that Eq. (2.37) holds only in a Euclidean universe. Let us define the curvature parameter

$$\Omega_{\rm K} = 1 - \Omega_0, \tag{2.38}$$

where $\Omega_0$ is the ratio of total to critical density today, including contributions from matter, radiation, and any other form of energy such as a cosmological constant (we will systematically introduce the $\Omega$ notation in Sec. 2.4). If the curvature is nonzero, $\Omega_{\rm K} \neq 0$, the angular diameter distance generalizes to

$$d_A = \frac{a}{H_0\sqrt{|\Omega_{\rm K}|}} \begin{cases} \sinh[\sqrt{\Omega_{\rm K}} H_0 \chi] & \Omega_{\rm K} > 0, \\ \sin[\sqrt{-\Omega_{\rm K}} H_0 \chi] & \Omega_{\rm K} < 0. \end{cases} \tag{2.39}$$

Note that both of these expressions reduce to the Euclidean case in the limit that the curvature density $\Omega_{\rm K}$ goes to zero. If $\Omega_{\rm K} > 0$, the universe is open and $d_A$ is larger than in the

Euclidean case; conversely, for $\Omega_K < 0$, a closed universe, $d_A$ is smaller than in the $\Omega_K = 0$ case (see Fig. 9.14 for an intuitive illustration of these trends).

Another way of inferring distances in astronomy is to measure the flux from an object of known luminosity ("standard candle;" with the discovery of gravitational-wave sources, we now also have "standard sirens," to which all of the following applies as well). For nearby objects, the flux $F$ observed at a distance $d$ from a source of known luminosity $L$ is

$$F = \frac{L}{4\pi d^2}, \tag{2.40}$$

since the total luminosity through a spherical shell with area $4\pi d^2$ is constant and equal to $L$. How does this result generalize to an expanding universe? Again it is simplest to work on the comoving grid, this time with the source centered at the origin. The flux we observe is

$$F = \frac{L(\chi)}{4\pi \chi^2(a)} \tag{2.41}$$

where $L(\chi)$ is the luminosity through a comoving spherical shell with radius $\chi(a)$. To further simplify, let us assume that the photons are all emitted with the same energy. Then $L(\chi)$ is this energy multiplied by the number of photons passing through a comoving spherical shell per unit time. In a fixed time interval, photons travel farther on the comoving grid at early times than at late times since the associated physical distance at early times is smaller. Therefore, the number of photons crossing a shell in a fixed time interval will be smaller today than at emission, smaller by a factor of $a$. Similarly, the energy of the photons will be smaller today than at emission, because of expansion. Therefore, the energy per unit time passing through a comoving shell at a distance $\chi(a)$ (i.e., our distance) from the source will be a factor of $a^2$ smaller than the luminosity at the source. The flux we observe therefore will be

$$F = \frac{La^2}{4\pi \chi^2(a)} \tag{2.42}$$

where $L$ is the luminosity at the source.[2] We can then use Eq. (2.40) to *define the luminosity distance* in a Euclidean expanding universe:

$$d_L^{\rm Euc} \equiv \frac{\chi}{a}. \tag{2.43}$$

The luminosity distance is also shown in Fig. 2.3. We thus have $d_L = d_A/a^2$, and this relation also holds in curved universes, so Eq. (2.39), when divided by $a^2$, similarly yields $d_L$ for $\Omega_K \neq 0$.

All three distances are larger in a universe with dark energy than in one without. This follows from the fact that dark energy leads to an expansion that is accelerating. Thus, for

[2] In general there is one more difference that needs to be accounted for: the observed luminosity is related to the emitted luminosity at a different wavelength. Here we have assumed a detector which counts all the photons.

a fixed expansion rate $H_0$ today, the universe was expanding more slowly in the past if dark energy is present. We have already seen in Ch. 1 that a universe with dark energy is older. This also means that light emitted at a given redshift (i.e. scale factor) had more time to travel, and hence covered a larger distance. Distant objects will therefore appear fainter to us than if the universe was dominated by matter only.

## 2.3 Evolution of energy

After having familiarized ourselves with the expanding spacetime, we now turn to the constituents of the universe. How do we characterize matter, radiation and other stuff in the universe? For now, we are again interested only in the smooth background universe, so only the mean quantities are of interest. There is no mean net momentum or velocity, since this would break the isotropy of the universe. Essentially, we are left with the mean density and pressure as the only properties of the various constituents that are relevant for the background universe. Just as the energy and momentum of a particle are combined into a relativistic 4-momentum, the energy density and pressure can be combined into a relativistic tensor, the *energy-momentum tensor*, which in the isotropic smooth universe assumes a very simple form:

$$T^{\mu}{}_{\nu} = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & \mathcal{P} & 0 & 0 \\ 0 & 0 & \mathcal{P} & 0 \\ 0 & 0 & 0 & \mathcal{P} \end{pmatrix} \tag{2.44}$$

where $\mathcal{P}$ is the pressure. This simple form is just a consequence of the symmetries of the FLRW metric. The fact that it appears to be the energy-momentum tensor of an ideal fluid at rest should not mislead us; several constituents in the universe in fact do not behave as fluids. The energy-momentum tensor is precisely the entity appearing on the right-hand side of the Einstein equations so deriving it will become a routine calculation for us.

How do the components of the energy-momentum tensor evolve with time? To gain intuition, consider first the case of a fluid in the absence of gravity, and when velocities are negligible. The pressure and energy density in that case evolve according to the continuity equation, $\partial\rho/\partial t = 0$, and the Euler equation, $\partial\mathcal{P}/\partial x^i = 0$. Can this also be promoted to a 4-component conservation equation for the energy-momentum tensor, perhaps $\partial T^{\mu}{}_{\nu}/\partial x^{\mu} = 0$? Almost: as explained in Box 2.2, coordinate derivatives of tensors have no meaning by themselves in general relativity, as they are coordinate-dependent. Instead, we have to use the *covariant* derivative:

$$\nabla_{\mu} T^{\mu}{}_{\nu} \equiv \frac{\partial T^{\mu}{}_{\nu}}{\partial x^{\mu}} + \Gamma^{\mu}{}_{\alpha\mu} T^{\alpha}{}_{\nu} - \Gamma^{\alpha}{}_{\nu\mu} T^{\mu}{}_{\alpha} = 0. \tag{2.45}$$

This is the general-relativistic version of the continuity and Euler equations, or, more generally, the statement of local energy and momentum conservation.

## 2.2 Tensors and derivatives

In curved space, the metric is important not only for calculating distances, but also for taking derivatives. As always, the starting point is that physical conclusions should be independent of the coordinates chosen. Consider first a simple function $\phi(x)$ defined on the curved space, which we call a *scalar*. When changing coordinates $x^\mu \to \hat{x}^\mu$, it transforms almost trivially as

$$\hat{\phi}(\hat{x}) = \phi(x[\hat{x}]). \tag{2.46}$$

That is, the function $\phi$ just depends on which physical point in space the coordinates $x$ or $\hat{x}$ point to in their respective coordinate system.

We can generate a covariant vector from a scalar $\phi(x)$ by using the partial derivative with respect to the coordinates:

$$A_\mu = \frac{\partial}{\partial x^\mu}\phi\,. \tag{2.47}$$

$A_\mu$ points in the direction where $\phi$ changes most rapidly, and one can hence think of this vector as an arrow. If we change coordinates, we now have, via the chain rule [cf. Eq. (2.14)]

$$\hat{A}_\mu = \frac{\partial}{\partial \hat{x}^\mu}\hat{\phi} = \frac{\partial x^\alpha}{\partial \hat{x}^\mu}\frac{\partial}{\partial x^\alpha}\phi = \frac{\partial x^\alpha}{\partial \hat{x}^\mu}A_\alpha\,. \tag{2.48}$$

This transformation law applies generally to all covariant (lower-index) vectors, regardless of whether they are given by a derivative of a scalar or not. More generally, for a two-index tensor such as the metric, we have

$$\hat{g}_{\mu\nu} = \frac{\partial x^\alpha}{\partial \hat{x}^\mu}\frac{\partial x^\beta}{\partial \hat{x}^\nu}g_{\alpha\beta}\,. \tag{2.49}$$

Now, can we take a derivative of a vector to obtain a tensor, such as $M_{\mu\nu} = \partial_\mu A_\nu$? The answer is no: this object does not obey the tensor transformation law, as can be verified by taking a derivative of Eq. (2.48). However, the metric allows us to construct a *covariant derivative* $\nabla_\mu$ such that $\nabla_\mu A_\nu$ *is* a tensor. Thus, whenever one takes a derivative of a vector or tensor in relativity to generate a new physical field, it should be the covariant derivative.

What spoils the transformation law for the ordinary derivative of a vector is the same term, $\partial^2 x^\alpha/\partial\hat{x}^\mu\partial\hat{x}^\nu$, that led us to introduce the Christoffel symbols in the geodesic equation in Sect. 2.1.2. Indeed, the Christoffel symbols are all we need to define the covariant derivative. You can verify that

$$\nabla_\mu A_\nu \equiv \partial_\mu A_\nu - \Gamma^\alpha{}_{\mu\nu}A_\alpha \tag{2.50}$$

satisfies the proper tensor transformation law Eq. (2.49). For this, our previous definition of Christoffel symbols as the term in brackets in Eq. (2.19), generalized to four dimensions, is useful. Similarly, for contravariant vectors we have (note the change in sign)

$$\nabla_\mu A^\nu \equiv \partial_\mu A^\nu + \Gamma^\nu{}_{\mu\alpha}A^\alpha. \tag{2.51}$$

For tensors, we have one Christoffel term for each index. For example,

$$\nabla_\mu T_\nu{}^\kappa = \partial_\mu T_\nu{}^\kappa - \Gamma^\lambda{}_{\mu\nu}T_\lambda{}^\kappa + \Gamma^\kappa{}_{\mu\lambda}T_\nu{}^\lambda. \tag{2.52}$$

As an example application of the covariant derivative, you can show, using Eq. (2.26), that the geodesic equation (2.20) can be written in the compact, manifestly covariant form

$$P^\alpha \nabla_\alpha P^\mu = 0. \tag{2.53}$$

Eq. (2.45) consists of four separate equations ($\nu = 0, 1, 2, 3$). Let us apply them to the case of a smooth expanding universe described by the FLRW metric, as captured by Eq. (2.44). We begin with the $\nu = 0$ component, which is

$$\frac{\partial T^\mu{}_0}{\partial x^\mu} + \Gamma^\mu{}_{\alpha\mu} T^\alpha{}_0 - \Gamma^\alpha{}_{0\mu} T^\mu{}_\alpha = 0\,. \tag{2.54}$$

Since we are assuming isotropy, $T^i{}_0$ vanishes, so the indices $\mu$ in the first term and $\alpha$ in the second must be equal to zero:

$$-\frac{\partial \rho}{\partial t} - \Gamma^\mu{}_{0\mu}\rho - \Gamma^\alpha{}_{0\mu} T^\mu{}_\alpha = 0. \tag{2.55}$$

From Eqs. (2.24)–(2.25), $\Gamma^\alpha{}_{0\mu}$ vanishes unless $\alpha$, $\mu$ are spatial indices equal to each other, in which case it is $\dot{a}/a$. So, the conservation law in an expanding universe reads

$$\frac{\partial \rho}{\partial t} + \frac{\dot{a}}{a}[3\rho + 3\mathcal{P}] = 0. \tag{2.56}$$

Rearranging terms, we have

$$a^{-3}\frac{\partial[\rho a^3]}{\partial t} = -3\frac{\dot{a}}{a}\mathcal{P}. \tag{2.57}$$

The conservation law can be applied immediately to glean information about the scaling of both matter and radiation with the expansion. Nonrelativistic matter has effectively zero pressure,[3] so

$$\frac{\partial[\rho_{\rm m} a^3]}{\partial t} = 0, \tag{2.58}$$

implying that the energy density of matter follows $\rho_{\rm m} \propto a^{-3}$. We anticipated this result in Ch. 1 based on the simple notion that the particle mass remains constant, while the number density scales as the inverse volume. The application to radiation also offers no surprises. Radiation has $\mathcal{P}_{\rm r} = \rho_{\rm r}/3$ (Exercise 2.9), so working from Eq. (2.56),

$$\begin{aligned}\frac{\partial \rho_{\rm r}}{\partial t} + \frac{\dot{a}}{a} 4\rho_{\rm r} &= a^{-4}\frac{\partial[\rho_{\rm r} a^4]}{\partial t} \\ &= 0.\end{aligned} \tag{2.59}$$

[3] Recall that we set $c = 1$ here; we are really comparing $\mathcal{P}$ with $\rho c^2$.

Therefore, the energy density of radiation $\rho_{\rm r} \propto a^{-4}$, accounting for the decrease in energy per particle as the universe expands.

We can summarize the cases of matter and radiation in one equation, and generalize the evolution results to other constituents, by defining the *equation of state* parameter $w_s$,

$$w_s \equiv \frac{\mathcal{P}_s}{\rho_s}, \tag{2.60}$$

where $s$ stands for any constituent of the universe. Matter corresponds to $w = 0$, radiation to $w = 1/3$, and, as we will see later, a cosmological constant has $w = -1$. However, the equation of state does not have to be time-independent in general. Eq. (2.56) can be integrated to find the evolution of any constituent $s$ with a time-dependent equation of state $w_s(a)$:

$$\rho_s(a) \propto \exp\left\{-3\int^a \frac{da'}{a'}[1 + w_s(a')]\right\} \overset{w_s=\text{const}}{\propto} a^{-3(1+w_s)}. \tag{2.61}$$

The second proportionality holds if $w_s$ is time-independent.

Let us briefly mention the $\nu = i$ part of Eq. (2.45). It turns out that in the smooth background universe, this equation is trivially zero. This follows from the fact that the spatial part of the metric is isotropic, since the universe is expanding equally in all directions, and that matter does not have any peculiar motions in the smooth universe. The Euler equation is naturally trivial if the velocity vanishes! Unsurprisingly, this will change once we begin to allow for structure in the universe.

So far, we have phrased the properties of different constituents in the universe in terms of their mean density and pressure. These are *macroscopic* quantities. Microscopically, within a volume centered on any given point, matter and radiation are composed of many interacting (or non-interacting) particles of different species. These can be described statistically by their *distribution functions*. Given an infinitesimal volume element $d^3x$ centered around point $\boldsymbol{x}$ and time $t$, the distribution function $f_s(\boldsymbol{x}, \boldsymbol{p}, t)$ counts the number of particles of a given species $s$ within an infinitesimal momentum-space element $d^3p$.[4] The total energy density of a given species is then obtained by summing the energy over all phase-space elements, weighted by the number of particles: $\sum f_s(\boldsymbol{x}, \boldsymbol{p}, t)E_s(p)$ with $E_s(p) = \sqrt{p^2 + m_s^2}$. How many phase-space elements are there in a region of 6D volume $d^3x d^3p$? By Heisenberg's principle, no particle can be localized into a region of phase space smaller than $(2\pi\hbar)^3$, so this is the size of a fundamental element. Therefore, the number of phase-space elements in $d^3x d^3p$ is $d^3x d^3p/(2\pi\hbar)^3$ (see Fig. 2.4). Dividing by the volume $d^3x$ yields the energy density in species $s$, which is

$$\rho_s(\boldsymbol{x}, t) = g_s \int \frac{d^3p}{(2\pi)^3} f_s(\boldsymbol{x}, \boldsymbol{p}, t)E_s(p) \tag{2.62}$$

[4] Recall that by $p$ here we mean not the comoving momentum $P$ defined in Eq. (2.26), but rather the physical momentum defined in Eq. (2.32).

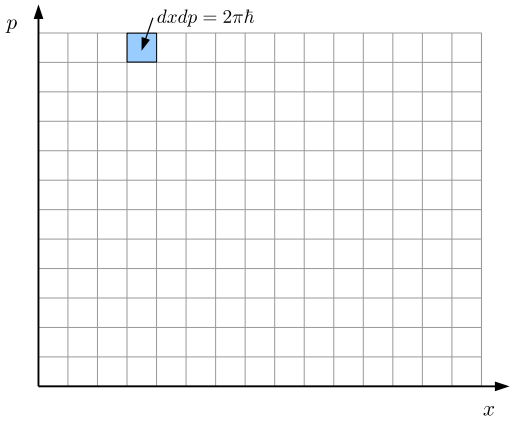



**FIGURE 2.4** Phase space of position and momentum in one dimension. The volume of each cell is $2\pi\hbar$, the smallest region into which a particle can be confined because of Heisenberg's principle. To count the appropriate number of cells, therefore, the phase-space integral in one dimension must be $\int dxdp/(2\pi\hbar)$.

where $g_s$ is the degeneracy of the species (e.g., equal to 2 for the photon for its two spin states), and we have gone back to $\hbar = 1$.

The macroscopic pressure within a given volume element corresponds to the force exerted per unit area on a fictitious boundary wall by the elastic collisions of the particles. For simplicity, let us consider $N$ nonrelativistic particles in a volume $V$. Then the pressure in the $x$ direction is given by

$$\mathcal{P} = \frac{N}{V} m v_x^2 = \frac{1}{3}\frac{N}{V} m|\boldsymbol{v}|^2, \tag{2.63}$$

where $v_x$ is the RMS velocity in the $x$-direction of the particles, and we have used the fact that the velocity dispersion and hence the pressure are isotropic. Thus, the pressure is given by $2/3$ of the sum over the *kinetic* energies of the particles within that volume. It can thus be similarly expressed as an integral over the distribution function, which after generalizing $m|\boldsymbol{v}|^2 \to p^2/E_s(p)$, which also holds for relativistic particles, becomes

$$\mathcal{P}_s(\boldsymbol{x}, t) = g_s \int \frac{d^3p}{(2\pi)^3} f_s(\boldsymbol{x}, \boldsymbol{p}, t) \frac{p^2}{3E_s(p)}. \tag{2.64}$$

Through most of the early universe, reactions proceeded rapidly enough to keep particles in equilibrium, with different species sharing a common temperature. We will often want to express the energy density and pressure in terms of this temperature. In equilibrium at temperature $T$, bosons, such as photons, have Bose–Einstein distributions, $f_s(\boldsymbol{x}, \boldsymbol{p}, t) = f_{\rm BE}(E_s(p))$, with

$$f_{\rm BE}(E) = \frac{1}{e^{(E-\mu)/T} - 1}, \tag{2.65}$$

and fermions, such as electrons, have Fermi–Dirac distributions,

$$f_{\rm FD}(E) = \frac{1}{e^{(E-\mu)/T}+1}, \tag{2.66}$$

with $\mu$ being the chemical potential. It should be noted that these equilibrium distributions do not depend on position $\boldsymbol{x}$ or on the direction of the momentum $\hat{\boldsymbol{p}}$, simply on the magnitude $p$ via $E_s(p)$.

For photons and neutrinos, the chemical potential is much smaller than the temperature. Photon number is not conserved (e.g., photons can be created and destroyed in double-Compton scattering, which is very efficient in the early universe), while for neutrinos, there is likely only a very small asymmetry between particles and anti-particles. In these cases, then, the distribution function depends only on $E/T$ and the pressure satisfies (Exercise 2.9)

$$\frac{\partial \mathcal{P}_s}{\partial T} = \frac{\rho_s + \mathcal{P}_s}{T}. \tag{2.67}$$

This relation can be used to show that the entropy density in the universe scales as $a^{-3}$. To see this, let us rewrite the continuity equation (2.57) as

$$a^{-3}\frac{\partial\left[(\rho+\mathcal{P})a^3\right]}{\partial t} - \frac{\partial \mathcal{P}}{\partial t} = 0. \tag{2.68}$$

In the background universe, we can rewrite the derivative of the pressure with respect to time in terms of the temperature as $(dT/dt)(\partial\mathcal{P}/\partial T)$, which also holds for multiple constituents. So,

$$\begin{aligned} a^{-3}\frac{\partial[(\rho+\mathcal{P})a^3]}{\partial t} - \frac{dT}{dt}\frac{\rho+\mathcal{P}}{T} &= a^{-3}T\frac{\partial}{\partial t}\left[\frac{(\rho+\mathcal{P})a^3}{T}\right] \\ &= 0. \end{aligned} \tag{2.69}$$

So the entropy density[5]

$$s \equiv \frac{\rho+\mathcal{P}}{T} \tag{2.70}$$

scales as $a^{-3}$. This scaling holds for the total entropy of all species in equilibrium (i.e. with the same temperature) as well as individual species in general. In fact, even if two species have different temperatures, the sum of their entropy densities still scales as $a^{-3}$. We will make use of this fact shortly when computing the relative temperatures of neutrinos and photons in the universe.

[5] Technically, there is another term in the entropy density—proportional to the chemical potential—but, as mentioned above, this term is usually irrelevant in cosmology. Even with nonzero chemical potential, though, the entropy density scales as $a^{-3}$.

## 2.4 Cosmic inventory

Armed with an expression for the energy density of a given particle species (Eq. (2.62)), and a knowledge of how it evolves in time (Eq. (2.57)), we can now tackle quantitatively the question of how much energy is contributed by the different constituents of the universe. Note that a constituent can be made up of several particle species (e.g., electrons and nuclei in the case of baryons), but each has to have the same equation of state (e.g., nonrelativistic or ultra-relativistic).

It will be useful to have all energy densities in the same units. The simplest way to do this is to divide all energy densities by the critical density today, Eq. (1.4), and define the *density parameters*[6]

$$\Omega_s \equiv \frac{\rho_s(t_0)}{\rho_{\rm cr}}, \tag{2.71}$$

where $s$ stands for any constituent of the universe: cold dark matter (c), baryons (b), photons ($\gamma$), neutrinos ($\nu$), and a cosmological constant ($\Lambda$) or dark energy. We will also use a subscript r for all radiation constituents (photons and ultra-relativistic neutrinos), and m for the total nonrelativistic matter: $\Omega_{\rm m} = \Omega_{\rm b} + \Omega_{\rm c}$.

Thus, we can write the density of constituent $s$ as a function of scale factor as

$$\rho_s(a) = \Omega_s \rho_{\rm cr} a^{-3(1+w_s)}, \tag{2.72}$$

assuming that its equation of state $w_s$ is time-independent. Now recall that $\rho_{\rm cr} = 3H_0^2/8\pi G$, and that $H_0$ is not perfectly known. This means that any precise constraint on the physical mean density of baryons $\rho_{\rm b}$, say, really constrains the parameter combination $\Omega_{\rm b}h^2$. For this reason, constraints are often phrased in terms of this combination of parameters in the literature, and it is even given its own symbol: $\omega_s \equiv \Omega_s h^2$; indeed, we have already encountered this combination in the label of the $x$-axis in Fig. 1.6.

### 2.4.1 Photons

The majority of the radiation contribution to the cosmic energy budget is in the form of the *cosmic microwave background* (CMB). Given its black-body nature, i.e. Bose–Einstein distribution function, the energy density associated with this radiation is

$$\rho_\gamma = 2\int \frac{d^3p}{(2\pi)^3}\frac{p}{e^{p/T}-1}. \tag{2.73}$$

The factor of 2 in front of Eq. (2.73) accounts for the two spin states of the photon. The energy of a given state is simply equal to $p$ since the photon is massless. The chemical potential is zero; we expect this theoretically because early in the universe, photon number is

[6] The critical density can be defined as a time-dependent quantity by replacing $H_0$ in Eq. (1.4) with $H(t)$. Correspondingly, time-dependent density parameters $\Omega_s(t)$ are sometimes used in the literature. In this book, apart from Ch. 12, we will always define $\rho_{\rm cr} \equiv \rho_{\rm cr,0}$ and $\Omega_s \equiv \Omega_{s,0}$ to be at today's epoch $t_0$, and refrain from adding a subscript 0.

not conserved. We also know it observationally because the spectrum of the CMB has been measured so accurately. The limits on a chemical potential are $\mu/T < 9 \times 10^{-5}$, as obtained from data of the FIRAS instrument aboard the COBE satellite (Fixsen et al., 1996), so $\mu$ can be safely ignored. Moreover, as we mentioned in Ch. 1, FIRAS measured the temperature of the CMB extraordinarily precisely: $T_0 = 2.726 \pm 0.001$ K (Fixsen, 2009). Since there is no angular dependence in the integrand of Eq. (2.73), the angular integral yields a factor of $4\pi$ and we are left with a one-dimensional integral. Define the integration variable $x \equiv p/T$: then

$$\rho_\gamma = \frac{8\pi T^4}{(2\pi)^3} \int_0^\infty dx \frac{x^3}{e^x - 1}. \tag{2.74}$$

The integral can be expressed in terms of the Riemann $\zeta$ function (Eq. (C.29)); it is $6\zeta(4) = \pi^4/15$, so that we finally have

$$\rho_\gamma = \frac{\pi^2}{15} T^4. \tag{2.75}$$

Since we derived that the energy density of radiation scales as $a^{-4}$ (Eq. (2.59)), the temperature of the CMB must scale as $a^{-1}$. In fact, with both $E \propto 1/a$ for each photon and $T \propto 1/a$, we see from Eq. (2.65) that the expansion of the universe preserves the equilibrium form of the distribution function.

We thus have for the photon density parameter today

$$\Omega_\gamma h^2 = 2.47 \times 10^{-5}. \tag{2.76}$$

To get this result, it is useful to remember the conversion between Kelvin and eV: 11605 K = 1 eV. So, photons make up a very small fraction of the universe's energy budget today. To reiterate an important point: $\rho_\gamma$ in Eq. (2.75) depends only on time. This is because we have used the zeroth-order Bose–Einstein distribution function for the photons. In fact there are small perturbations around this zeroth-order distribution function. These do have a spatial and momentum dependence and correspond to the anisotropies in the CMB.

### 2.4.2 Baryons

Following standard conventions in cosmology, we refer to all ordinary matter, i.e. nuclei and electrons, as *baryons*, even though this is technically incorrect as electrons are leptons. However, nuclei are so much more massive than electrons that virtually all of the mass *is* in the baryons. Unlike the CMB, baryons cannot be simply described with an equilibrium distribution function. This is because baryons come in many different phases: diffuse neutral gas and ionized plasma, stars and planets, compact objects, and so on. This makes a baryonic inventory much more difficult.

There have been many attempts at a direct count of baryons in the past (Fukugita et al., 1998; Shull et al., 2012) One approach is to count the amount of baryons in stars and diffuse gas in galaxies and groups of galaxies, although hot ionized gas (with temperatures less than a keV) is difficult to detect, making such estimates uncertain. A second way to count

baryons is by looking at the spectra of distant quasars, extremely bright active galactic nuclei. The amount of light absorbed from these beacons is a measure of the intervening hydrogen, and hence the baryon density. However, the amount of mean absorption per hydrogen atom depends on the thermal state of the intergalactic medium, which is not very well known.

Let us thus focus on measurements in the early universe, which rely on straightforward nuclear and atomic physics: Big Bang Nucleosynthesis (BBN) and the CMB. The abundance of light elements formed during BBN (Sect. 1.3) depends on the total physical baryon density in the universe, and thus constrains $\Omega_{\rm b}h^2$ (see Fig. 1.6 and the discussion after Eq. (2.72)). We will see how this works in Ch. 4. Deuterium is most sensitive to the baryon density, and measurements of the fractional amount of deuterium in high-redshift absorption systems combined with BBN yield $\Omega_{\rm b}h^2 = 0.0222 \pm 0.0005$ (Cooke et al., 2018).

The baryon density also affects the plasma oscillations in the early universe, whose imprints we see as anisotropies in the CMB, as shown in Fig. 1.10. We will derive the details in Ch. 9. The Planck team constrained the baryon density to be $\Omega_{\rm b}h^2 = 0.0225 \pm 0.0003$ (Planck Collaboration, 2018b), a constraint which is only weakly dependent on the assumed cosmological model.

Given that current best estimates of the Hubble constant give $h \simeq 0.7$, these two spectacularly precise constraints agree on a baryon density $\Omega_{\rm b}$ that is roughly 5% of the critical density today. Notice that these early-universe measurements also include any baryonic matter that would be very difficult to count in today's universe, for example if it had collapsed into quiescent black holes or neutron stars. Nevertheless, the astrophysical constraints are also largely in agreement with this value given the estimated uncertainties. Finally, the total matter density in the universe is larger than this by a factor $\sim 6$, so more than 80% of the matter in the universe has to be non-baryonic.

### 2.4.3 Dark matter

As we mentioned in Ch. 1, the overwhelming evidence for (non-baryonic) dark matter is not a new revelation to astronomers, who have found corresponding evidence within our Milky Way and local group, as well as other galaxies and clusters of galaxies. But how do we measure the total density of matter? Unlike for baryons, we cannot use nuclear and atomic physics, but have to rely on gravity.

The anisotropies in the CMB (Ch. 9) provide a measurement of the physical matter density parameter $\Omega_{\rm m}h^2$. The sensitivity of the CMB to the matter density is both due to the effect of matter on the expansion history in the early universe, as well as the fact that dark matter dominates the gravitational potential wells which also leave their imprint in the CMB anisotropies. Assuming the concordance model, the Planck team reported $\Omega_{\rm m}h^2 = 0.1431 \pm 0.0025$ (Planck Collaboration, 2018b). Therefore, again invoking our knowledge of the Hubble constant, the CMB observations are consistent with a matter density equal to about 30% of the critical density.

The distance-redshift relation in the late universe, as probed by standard candles and rulers, constrains $\Omega_{\rm m}$ alone. When combined with the CMB, the constraint becomes very tight, yielding $\Omega_{\rm m} = 0.311 \pm 0.006$.

As we will see in Ch. 11 and Ch. 13, large-scale structure provides two beautiful ways to probe gravitational potential wells and hence the amount of matter: galaxy velocities and gravitational lensing. Velocities are probed through the characteristic distortion they imprint on the three-dimensional statistics of galaxy number counts. Gravitational lensing is detected through the statistics of galaxy *shapes*. As an example, measurements of weak gravitational lensing and galaxy clustering using the first year of data from the Dark Energy Survey resulted in a constraint of $\Omega_{\rm m} = 0.27^{+0.03}_{-0.02}$ (Abbott et al., 2018). The slight discrepancy between this number and those driven mostly by the CMB is useful to point out (even though both may have changed slightly by the time you read this) because it $(i)$ highlights the robust conclusion from all probes that the total matter density is roughly 30% of the critical density and $(ii)$ acknowledges that at any given time, there are often hints of tension in the values of parameters inferred from different probes. Whether these are simply statistical fluctuations that will go away with more data, or indicate profound cracks in the concordance model, is one of the exciting open questions in modern cosmology.

Finally, another way of measuring the total mass density is to pick out observations sensitive to $\Omega_{\rm b}/\Omega_{\rm m}$ and use the value of $\Omega_{\rm b}$, determined through either BBN or CMB, to infer the matter density. Massive galaxy clusters are perhaps the most promising target, since most of the baryonic mass in a galaxy cluster is in the form of hot gas which is observable through its thermal X-ray emission or the so-called Sunyaev–Zel'dovich (SZ) effect (see Sect. 12.5 and Sect. 11.3, respectively). If this ratio is characteristic of the universe as a whole—it probably is to a good approximation—then the cosmic baryon to matter ratio is $\Omega_{\rm b}/\Omega_{\rm m} = (0.089 \pm 0.012)h^{-3/2}$ (Mantz et al., 2014). Since baryons make up about 5% of the critical density, the total matter density is inferred again to be roughly 30% of the critical density.

We conclude that there is now agreement among a wide variety of probes that the total matter density in the universe is about 30% of the critical density, with 80% of that being in the form of non-baryonic dark matter.

### 2.4.4 Neutrinos

The next constituent we need to consider are neutrinos. Unlike photons and baryons, cosmic neutrinos have not been observed directly, so arguments about their contribution to the energy density are necessarily theoretical. However, these theoretical arguments are quite strong, based on very well-understood physics. Moreover, the CMB anisotropies constrain the total density in relativistic particles $\Omega_{\rm r}h^2$ in the early universe. Experiments such as the Planck satellite have found clear evidence for an amount of relativistic particles (in addition to the known photons) that is consistent with the expected neutrino contribution.

Let us sum up what we know about these particles:

- There are three generations of neutrinos[7].
- There is one spin degree of freedom each for the neutrino and antineutrino of each generation.
- Neutrinos are fermions and follow the Fermi–Dirac distribution function when in equilibrium.

We can use this information to evaluate the energy density of neutrinos in the universe, relating it to the photon energy density $\rho_\gamma$ for convenience. The first two items on the list imply that the degeneracy factor of neutrinos is equal to 6. The third means we need to change the denominator in the integrand in Eq. (2.73) to $e^{p/T}+1$. The resulting Fermi–Dirac energy integral is smaller by a factor of 7/8 compared to the corresponding Bose–Einstein integral. Finally, since the energy density of a massless particle scales as $T^4$, we can write

$$\rho_\nu = 3 \times \frac{7}{8} \times \left(\frac{T_\nu}{T}\right)^4 \rho_\gamma . \tag{2.77}$$

We then only need to determine $T_\nu$, which, as you might have guessed from how we wrote this equation, is different from the photon temperature $T$.

For this, let us first consider the production of neutrinos in the early universe. A basic understanding of the interaction rates of neutrinos (Fig. 1.4) enables us to argue that neutrinos were once kept in equilibrium with the rest of the cosmic plasma. At later times, they lost contact with the plasma because their interactions are *weak*. The tricky part in determining the neutrino temperature is the annihilation of electrons and positrons when the cosmic temperature was of order the electron mass. Neutrinos lost contact with the cosmic plasma slightly before this annihilation, so they inherited almost none of the associated energy. The photons, which acquired the vast majority of it, are therefore hotter than the neutrinos.

We can account for the annihilation of electrons and positrons by using the fact that the total entropy density $s$ (Eq. (2.70)) scales[8] as $a^{-3}$. Massless bosons contribute $2\pi^2 T^3/45$ to the entropy density for each spin state, while massless fermions contribute 7/8 of this, and particles whose masses are larger than the temperature at that time contribute negligibly (Exercise 2.11). Before $e^+e^-$-annihilation, the relevant fermions are electrons and positrons (two spin states each), and (anti-)neutrinos (six, as we counted above). The bosons are photons (two spin states). So at $a = a_1$ before annihilation,

$$s(a_1) = \frac{2\pi^2}{45} T_1^3 \left[2 + \frac{7}{8}(4+6)\right]$$

[7] There is also the possibility that other types of neutrinos exist. These would have no interactions with the rest of the Standard Model particles, so are called *sterile* neutrinos. Even if they do exist, in many models their interactions with ordinary neutrinos are extremely weak and they have a very small cosmic abundance. Therefore, we neglect them here.

[8] As mentioned in the footnote in Sect. 2.3, this works regardless of our imperfect knowledge of the neutrino chemical potential.

$$= \frac{43\pi^2}{90} T_1^3 \tag{2.78}$$

where $T_1$ is the common temperature at $a_1$. After annihilation, at $a_2$, the electrons and positrons have gone away and the photon and neutrino temperatures are no longer identical: we must distinguish between them. On the other hand, all other remaining particles are much more massive and contribute negligibly. Therefore, the entropy density is

$$s(a_2) = \frac{2\pi^2}{45}\left[2T^3 + \frac{7}{8}6T_\nu^3\right]. \tag{2.79}$$

Equating $s(a_1)a_1^3$ with $s(a_2)a_2^3$ leads to

$$\frac{43}{2}(a_1 T_1)^3 = 4\left[\left(\frac{T}{T_\nu}\right)^3 + \frac{21}{8}\right]_{a_2} (T_\nu(a_2)a_2)^3. \tag{2.80}$$

But, neglecting the very small amount of energy received from $e^\pm$, the neutrino temperature scales as $a^{-1}$ throughout, so $a_2 T_\nu(a_2) = a_1 T_1$. Therefore, the ratio of the two temperatures is

$$\frac{T_\nu}{T} = \left(\frac{4}{11}\right)^{1/3}, \tag{2.81}$$

which continues to hold up until today. So, using Eq. (2.77), we obtain the neutrino density as

$$\rho_\nu = 3 \times \frac{7}{8} \times \left(\frac{4}{11}\right)^{4/3} \rho_\gamma. \tag{2.82}$$

It is tempting to use this result together with Eq. (2.76) to infer that the energy density of neutrinos today is $\Omega_\nu h^2 = 1.68 \times 10^{-5}$. However, observations in 1998 of neutrino oscillations (Fukuda et al., 1998) proved that neutrinos are not massless, as already hinted at from earlier observations of solar neutrinos (Bahcall, 1989). These measurements imply that the sum of masses of neutrinos of all generations has to be at least 0.06 eV.[9] In the early universe (up until recombination), the neutrino masses are indeed entirely negligible, so that Eq. (2.82) holds. However, later on neutrinos transitioned from relativistic to nonrelativistic as the temperature dropped beneath their mass. The same happened to other massive particle species as well, but unlike other species, this transition happened relatively recently for neutrinos which leads to interesting effects on the growth of structure.

Since neutrinos have mass, we need to go further to determine their energy density today. For a single neutrino generation with mass $m_{\nu_i}$, the energy density is

$$\rho_{\nu_i} = 2\int \frac{d^3p}{(2\pi)^3} \frac{1}{e^{p/T_\nu} + 1} \sqrt{p^2 + m_{\nu_i}^2}. \tag{2.83}$$

[9] The oscillation experiments are sensitive to mass *differences*, $m_2^2 - m_1^2$, so the actual constraint is that the mass squared difference is of order $10^{-3}$ eV$^2$.

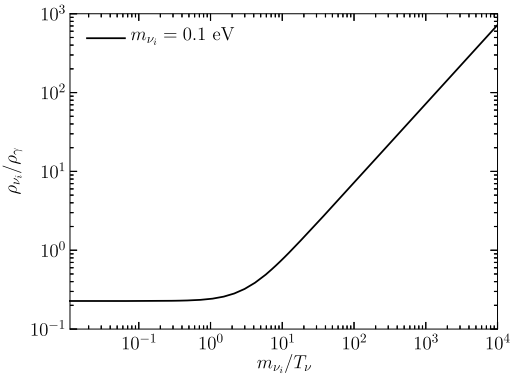



**FIGURE 2.5** Energy density of one generation of massive neutrinos relative to the energy density of the photons. At high temperatures, the ratio is a fixed constant; at low temperatures, the neutrino behaves like nonrelativistic matter (scaling as $a^{-3}$) and so begins to dominate over the photon density (which scales as $a^{-4}$).

This is similar to our expression for the photon density except that the nonzero mass is included and the distribution function is of Fermi–Dirac form. Note that the distribution is *not* strictly speaking a Fermi–Dirac distribution because the argument of the exponential is $p$, not the energy. This follows from the fact that after neutrinos decoupled they were no longer kept in equilibrium by scattering processes. Rather, they simply maintained their initial distribution (which was determined when the mass was much smaller than the temperature and therefore irrelevant) with the particle momenta redshifting as the universe expands; you can derive this in Exercise 3.9. At high temperatures, Eq. (2.83) reduces to one third of Eq. (2.82), so when considering neutrinos in the early universe, it is often sufficient to use Eq. (2.82). Indeed, we will do this shortly when we come to estimate the epoch at which the energy density of matter equals that of radiation. At sufficiently late times, when $T_\nu \ll m_{\nu_i}$, the energy density for a single massive neutrino is $m_{\nu_i} n_{\nu_i}$, just like for baryons and dark matter, with the neutrino number density $n_{\nu_i} = 3n_\gamma/11$ for a single generation (Exercise 2.12). As can be seen from Fig. 2.5, the transition takes place when $T_\nu \sim m_{\nu_i}$. Numerically, we obtain for the total neutrino density parameter today

$$\Omega_\nu h^2 = \frac{\sum_i m_{\nu_i}}{94 \text{ eV}} \tag{2.84}$$

where the sum is over the masses of all three neutrinos.

Those who trafficked in both astrophysics and particle physics (Gershtein and Zel'dovich, 1966; Szalay and Marx, 1976; Cowsik and McClelland, 1972) early on noted that the simple observation that the total density was not much greater than the critical density leads to constraints on neutrino mass that are much more stringent than those that were then obtainable from laboratory experiments. When the need for non-baryonic dark matter first became evident, a number of cosmologists (e.g., Gunn et al., 1978) proposed neutrinos as the natural candidate. Subsequent studies (Bond et al., 1980; White et al., 1983) of the structure of the universe with neutrinos as the dominant dark matter constituent showed

features that differ significantly from the actual universe (see Ch. 12). Nonetheless, the possibility that neutrinos might make up a *fraction* of the total density reemerged in the 1990s. We can then hope to detect a trace amount of neutrinos—corresponding to masses smaller than an eV—by observing its effect on large-scale structure (Ch. 8).

### 2.4.5 Epoch of matter–radiation equality

The epoch at which the energy density in matter equals that in radiation is called *matter–radiation equality*. It has a special significance for the generation of large-scale structure and for the development of CMB anisotropies, because perturbations grow at different rates in the two different eras (note that for large-scale structure, there is a third era: that of dark energy domination today; see Exercise 2.14). It is therefore a useful exercise to calculate the epoch of matter–radiation equality. To do this, we need to compute the energy density of both matter and radiation, and then find the value of the scale factor at which they were equal.

Using Eq. (2.76) and Eq. (2.82), we see that, as long as $T_\nu$ is much larger than all neutrino masses, the total energy density in radiation is

$$\frac{\rho_{\rm r}}{\rho_{\rm cr}} = \frac{4.15\times10^{-5}}{h^2a^4} \equiv \frac{\Omega_{\rm r}}{a^4}. \tag{2.85}$$

To calculate the epoch of matter–radiation equality, we equate Eqs. (2.85) and (2.72) to find

$$a_{\rm eq} = \frac{4.15\times10^{-5}}{\Omega_{\rm m}h^2}. \tag{2.86}$$

A different way to express this epoch is in terms of redshift $z$; the redshift of equality is

$$1+z_{\rm eq} = 2.38\times10^4\Omega_{\rm m}h^2. \tag{2.87}$$

Note that, as the amount of matter in the universe today, $\Omega_{\rm m}h^2$, goes up, the redshift of equality also goes up.

### 2.4.6 Dark energy

We now know that there is an additional ingredient in the universe's energy budget, *dark energy*, a substance whose equation of state $w$ is neither 0 (as it would be if the substance was nonrelativistic) or 1/3 (ultra-relativistic), but rather close to $-1$. A multitude of independent pieces of evidence has accumulated for the existence of dark energy, a substance that has this negative equation of state and does not participate in gravitational collapse. For one, we have strong evidence that the universe is Euclidean, with total density parameter close to 1. Since $\Omega_{\rm m} = 0.3$ is very far from 1 (and radiation is totally negligible today), something that does not clump as does matter has to make up this budgetary shortfall. Second, the expansion of the universe is accelerating, as measured by standard candles and rulers. As we will see in Ch. 3, accelerated expansion ($\ddot{a} > 0$) occurs only if the dominant constituent in the universe has a *negative* equation of state, i.e. negative pressure.

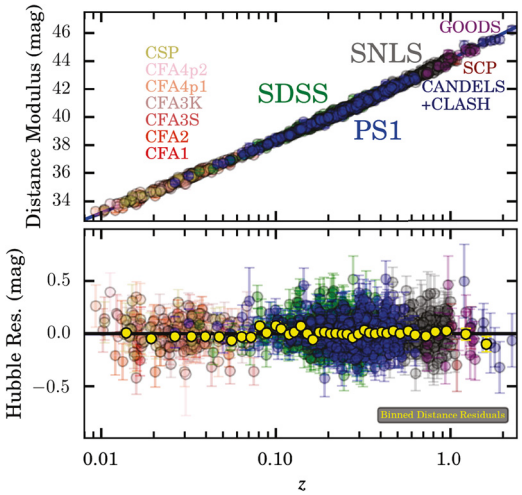



**FIGURE 2.6** Hubble diagram from the Pantheon sample of Type Ia supernovae. The top panel shows the distance modulus $m - M$, the difference of apparent and standardized absolute magnitudes of each Supernova (Eq. (2.88)), vs. redshift. The bottom panel shows the residuals relative to the prediction of the best-fitting Euclidean $\Lambda$CDM cosmology. The data unequivocally require the presence of dark energy. From Scolnic et al. (2018).

Evidence that $\Omega_{\rm m} \simeq 0.3$ has been accumulating since about 1980, and theoretical arguments that the total density is equal to the critical density are tied to inflation, which was proposed around the same time. The latter claims were bolstered by observations of the CMB in the late 1990s (Ch. 9). Around the same time, two groups (Riess et al., 1998, Perlmutter et al., 1999) observing supernovae reported direct evidence for an accelerating universe, one that is best explained by postulating the existence of dark energy. The evidence is based on measurements of the luminosity distance. As discussed in Sect. 2.2, the luminosity distance depends on the how rapidly the universe expanded in the past: $d_L \propto \int dz/H(z)$. An accelerating universe, one in which the expansion rate was lower in the past, would therefore have larger luminosity distances, and therefore standard candles like supernovae would appear fainter.

More concretely, the luminosity distance of Eq. (2.43) can be used to find the apparent magnitude $m$ of a source with absolute magnitude $M$. Magnitudes are related to fluxes and luminosities via $m = -(5/2)\log(F) + \text{constant}$ and $M = -(5/2)\log(L) + \text{constant}$. Since the flux scales as $d_L^{-2}$, the apparent magnitude $m = M + 5\log(d_L) + \text{constant}$. The convention is that

$$m - M = 5\log\left(\frac{d_L}{10\text{pc}}\right) + K \tag{2.88}$$

where $K$ is a correction ("$K$-correction") for the shifting of the spectrum into or out of the observed wavelength range due to expansion. $m - M$ is referred to as *distance modulus*.

The two groups in 1998 measured the apparent magnitudes of dozens of Type Ia supernovae, which are known to be standardizable candles, i.e., they have absolute magnitudes

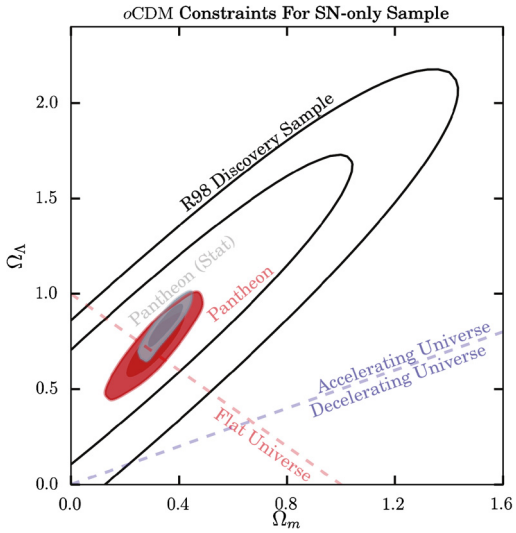



**FIGURE 2.7** Constraint on the parameters $(\Omega_\mathrm{m}, \Omega_\Lambda)$ quantifying the contribution of matter and cosmological constant to the cosmic energy budget from the Type Ia supernovae whose distance-redshift diagram is shown in Fig. 2.6 (filled contours). Here, a curved universe is allowed, with parameters corresponding to a Euclidean (or "flat") universe indicated by the line. The constraints from SNe clearly require an accelerating universe and that $\Omega_\Lambda > 0$. Also shown are the constraint contours of one of the original discovery papers (Riess et al., 1998). From Scolnic et al. (2018).

that can be determined from other observables, in particular the characteristic time it takes for the luminosity to decay after the peak. In practice, this is quite an involved analysis, requiring precise photometry and calibration. A more recent version of the result, a diagram of distance modulus vs. redshift, is shown in Fig. 2.6. By carefully accounting for statistical and systematic errors in the distance estimation, one can then obtain the best-fit parameters describing the expansion history of the universe. The result is shown in Fig. 2.7, based on the assumption that dark energy is a cosmological constant, but *not* restricted to a Euclidean universe. The two free cosmological parameters are then the matter density parameter $\Omega_\mathrm{m}$ and the corresponding parameter $\Omega_\Lambda$ for the cosmological constant. A universe with $\Lambda = 0$ (and hence $\Omega_\Lambda = 0$) is not compatible with observations. Instead, supernovae point to the concordance value of $\Omega_\Lambda \simeq 0.7$. Fig. 2.7 also shows the parameter constraints obtained by one of the original discovery analyses. Clearly, the supernova measurements have improved significantly since then.

Moreover, we now have yet another piece of independent evidence for dark energy: the Baryon Acoustic Oscillation (BAO) standard ruler (shown in Fig. 1.9) provides both a measurement of the angular diameter distance to a given redshift and the distance interval corresponding to a certain redshift interval. This is the derivative of the comoving distance with respect to redshift, $d\chi/dz = 1/H(z)$; we will see in Ch. 11 precisely how the BAO measurements lead to these constraints. A compilation of these measurements is shown in Fig. 2.8, along with the prediction of the Euclidean concordance cosmology including

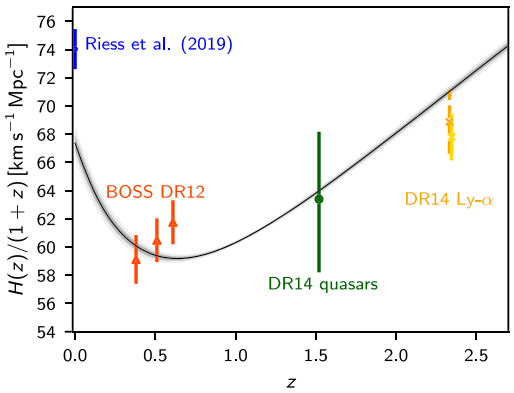



**FIGURE 2.8** Measurements of $\dot{a} = aH = H(z)/(1+z)$ from standard candles (data point at $z \simeq 0$) and the BAO standard ruler in the galaxy distribution (points at higher redshift). The line shows the best-fitting Euclidean $\Lambda$CDM model to the CMB and BAO measurements at $z < 1$. There is direct evidence that the comoving expansion rate decreases at higher redshifts (as expected if the universe is dominated by matter) but then increases again at low redshifts (requiring the existence of dark energy). From Planck Collaboration (2018b).

a cosmological constant. What is striking about this measurement is not merely that it allows for independent constraints on dark energy; it beautifully shows us *directly* that the expansion of the universe is accelerating. In a universe with matter and radiation (or any constituent with vanishing or positive pressure), the quantity $H(z)/(1+z) = aH = \dot{a}$ is monotonically decreasing. Indeed, the higher-redshift data points in Fig. 2.8 show this. However, we see that $aH$ has to *increase* in order to meet up with the locally measured Hubble rate. So, standard candles and rulers now allow us to see the presence of dark energy directly.

The existence of dark energy can be inferred not only using probes that measure the expansion history directly (sometimes called *geometric* probes). The accelerated expansion also directly affects the evolution of structure in the universe. We will see how this happens in Ch. 9, and discuss observable ramifications in Ch. 11. Growth of structure probes independently support the Euclidean concordance cosmology with $\Omega_\Lambda \simeq 0.7$. A compelling argument for the existence of dark energy is that both geometric (background) and dynamic (structure) probes agree on the same cosmological model.

So far, we have always talked about the cosmological constant $\Lambda$, with the one free parameter being the energy density associated with it, $\Omega_\Lambda$ in Fig. 2.7. However, this is only the simplest possibility for what dark energy could be, and introducing a constant carries its own set of problems (see Exercise 1.5). This is why we use "dark energy" rather than cosmological constant as a moniker. One generalization is to make the constant dynamical, turning the energy density associated with $\Lambda$ into the potential energy of a scalar field $V(\phi)$. This possibility is often referred to as *quintessence*. Yet another possibility is to modify general relativity itself, so that the acceleration is due to a modified behavior of gravity. For the reader interested in an overview of the model landscape, Mortonson et al. (2014) and Frieman et al. (2008) provide brief and comprehensive treatments of dark energy, re-

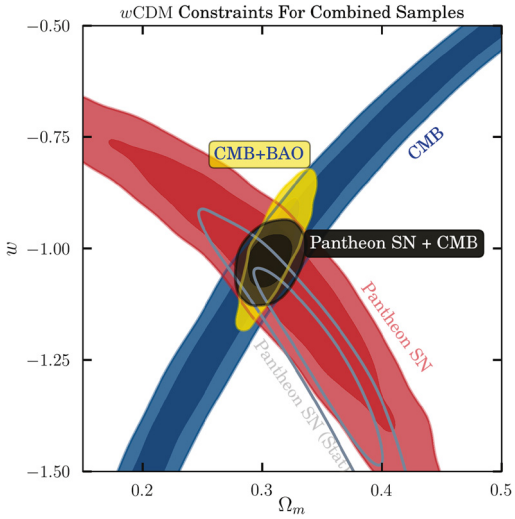



**FIGURE 2.9** Constraints, assuming a Euclidean universe, placed by different probes on the matter density $(\Omega_m)$ and constant equation of state of the dark energy $w = w_{\rm DE}$. A cosmological constant corresponds to $w = -1$. The constraints from supernovae, the BAO standard ruler, as well as the CMB all point towards a concordance model with $w_{\rm DE}$ close to $-1$. From Scolnic et al. (2018).

spectively; Joyce et al. (2016) and Clifton et al. (2012) do the same for modified gravity. Most pressing for us is the question of how we can distinguish among these possibilities given the data. Do we have to laboriously repeat the analysis of supernovae, BAO, and so on for each model of dark energy?

Fortunately not: as we argued at the beginning of Sect. 2.3, the form Eq. (2.44) of the energy-momentum tensor is completely general and is dictated by the symmetries of the FLRW spacetime. Hence, defining pressure via the equation of state $w_{\rm DE}(a)$, and given the continuity equation (2.57), whose solution is Eq. (2.61), the effect of a general dark energy on the expansion history is completely determined by the function $w_{\rm DE}(a)$.[10] The cosmological constant, as we will see in Sect. 3.1, simply adds a term $\Lambda\delta^{\mu}{}_{\nu}$ to the Einstein equations (when written with one upper index). Comparing this with Eq. (2.44) shows that the cosmological constant effectively has an energy-momentum tensor that is of perfect fluid form, with $\mathcal{P} = -\rho \propto \Lambda$ which implies an equation of state of $w_\Lambda = -1$. For a dynamical dark energy (e.g. quintessence), $w_{\rm DE} \geq -1$ (but still significantly below 0). Measuring the dark energy density as a function of cosmic time (i.e. at different redshifts) then allows us to constrain $w_{\rm DE}$ and hence distinguish between different dark energy scenarios.

Fig. 2.9 shows a current example of constraints on $w_{\rm DE}$, assuming a Euclidean universe. This figure drives home two points. First, so far all measurements are consistent with a cosmological constant; models with values of $w_{\rm DE}$ very different from $-1$ are ruled out.

[10] If general relativity is modified, we have to be a bit careful here. Nevertheless, one can always derive an equation of state dark energy *would have to have* in general relativity in order to produce the expansion history of a given modified gravity model.

Hence, $\Lambda$ remains our default scenario, resulting in the concordance $\Lambda$CDM model of cosmology. Second, different probes constrain the two parameters in different strengths, i.e. they have different *parameter degeneracies*, which also applies to parameters that are being left to vary but are not shown in Fig. 2.9, such as $H_0$. This means that it is highly beneficial to combine different cosmological probes. Over the past two decades, cosmologists have learned to appreciate that the power of the total combined measurements truly is more than the sum of the parts.

## 2.5 Summary

The smooth universe is described by the Friedmann–Lemaître–Robertson–Walker metric given in Eq. (2.12), which implies that physical distances are related to coordinate (comoving) distances with the time-dependent scale factor $a(t)$. By deriving the geodesic equation in this metric, we found that the physical momentum of particles decays as $1/a$. For massless particles like photons, this means that their energy redshifts as $1/a$ as well.

Measuring distances in the expanding universe is tricky, but all relevant distances can be obtained from the comoving distance between us and a source at redshift $z$:

$$\chi(z) = \int_0^z \frac{dz'}{H(z')}. \tag{2.89}$$

Another important distance is that which light could have traveled since $t = 0$. This, also called the conformal time, is

$$\eta = \int_0^t \frac{dt'}{a(t')} = \int_z^\infty \frac{dz'}{H(z')}. \tag{2.90}$$

The two quantities are simply related by

$$\chi(z) = \eta_0 - \eta(z), \tag{2.91}$$

where $\eta_0 \equiv \eta(z = 0)$. The conformal time will be the natural time variable when we come to consider the evolution of perturbations in the universe.

Photons in the universe have a Bose–Einstein distribution with zero chemical potential, so their energy density can be determined by measuring their temperature. Neutrinos are almost as abundant as photons, but there is some uncertainty in their energy density because of our ignorance of the neutrino masses (there is a lower limit of 0.06 eV for the sum of their masses and an upper limit of about 1 eV). In the early universe, this uncertainty is irrelevant since the temperatures are so much larger than the neutrino masses, so neutrinos behave relativistically. Thus, the uncertainty in neutrino mass affects neither Big Bang Nucleosynthesis at temperatures of order 1 MeV nor the epoch of matter–radiation equality at temperatures of order 1 eV. The neutrino temperature is a factor of $(4/11)^{1/3}$ smaller than the photon temperature. This, and the difference in statistics, implies that a single generation of massless neutrino has an energy density equal to 0.23 times that of

photons at those early times. Once $T_\nu \ll m_{\nu_i}$, this neutrino generation with mass $m_{\nu_i}$ contributes to $\Omega_\nu h^2$ with $m_{\nu_i}/94$ eV. In addition to photons and neutrinos, the universe consists of baryons ($\Omega_\mathrm{b} \simeq 0.05$); dark matter ($\Omega_\mathrm{c} \simeq 0.25$); and dark energy ($\Omega_\mathrm{DE} \simeq 0.7$), a constituent with negative pressure that, so far, is consistent with being a cosmological constant.

There is significantly more energy today in nonrelativistic matter and dark energy than in radiation. However, since the energy density of radiation scales as $a^{-4}$ while that of matter as $a^{-3}$, the very early universe was radiation dominated. The *equality* epoch at which the matter density was equal to the radiation density delineates these two regimes: $a_\mathrm{eq} = 4.15 \times 10^{-5}/\Omega_\mathrm{m} h^2$.

## Exercises

**2.1** Convert the following quantities by inserting the appropriate factors of $c$, $\hbar$, and $k_\mathrm{B}$:
- **(a)** $T_0 = 2.726\ \mathrm{K} \to \mathrm{eV}$
- **(b)** $\rho_\gamma = \pi^2 T_0^4/15 \to \mathrm{eV}^4$ and $\mathrm{g\,cm^{-3}}$
- **(c)** $1/H_0 \to \mathrm{cm}$
- **(d)** $m_\mathrm{Pl} \equiv \sqrt{\hbar c/G} = 1.2 \times 10^{19}\ \mathrm{GeV} \to \mathrm{K},\ \mathrm{cm^{-1}}, \mathrm{s^{-1}}$

**2.2** Show that the geodesic equation gets the correct equations of motion for a particle traveling freely in two dimensions using polar coordinates. You can get the Christoffel symbols one of two ways (or both!) and then proceed to **(b)**.
- **(a)** Get the Christoffel symbol either directly from the term in brackets in Eq. (2.19) or from the 2D metric

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \tag{2.92}$$

using Eq. (2.21). Show that the only nonzero Christoffel symbols are

$$\Gamma^2{}_{12} = \Gamma^2{}_{21} = \frac{1}{r} \quad ; \quad \Gamma^1{}_{22} = -r \tag{2.93}$$

with 1, 2 corresponding to $r$, $\theta$.
- **(b)** Write down the two components of the geodesic equation using these Christoffel symbols. Show that these give the correct equations of motion for a particle traveling in a plane.

**2.3** Find how the energy of a *massive*, nonrelativistic particle changes as the universe expands. Recall that in the massless case we used the fact that $g_{\mu\nu} P^\mu P^\nu = 0$. In this case, we have $g_{\mu\nu} P^\mu P^\nu = -m^2$.

**2.4** Show that the geodesic equation we derived in a Euclidean universe implies that, for massless particles,

$$\frac{d^2 \boldsymbol{x}}{d\eta^2} = 0 \tag{2.94}$$

where $\eta$ is the conformal time. This very important result says that, in comoving coordinates $(\eta, \boldsymbol{x})$, photons travel on straight lines as in Minkowski space.

**2.5** At early times, the cosmological constant can be neglected. Using this approximation, integrate Eq. (1.3) in a Euclidean universe to obtain $a(t)$. Using $T(t) = T_0/a(t)$, determine the times when the cosmic temperature was 0.1 MeV and 1/4 eV. We will see in Ch. 4 that these were the temperatures during two crucial epochs: Big Bang Nucleosynthesis and recombination.

**2.6** Derive some simple expressions for the conformal time $\eta$ as a function of $a$.

**(a)** Show that $\eta \propto a^{1/2}$ in a matter-dominated universe and $\eta \propto a$ in one dominated by radiation.

**(b)** Consider a Euclidean concordance universe with equality at $a_{\rm eq}$, at early times so that the effect of $\Lambda$ can be neglected. Show that

$$\eta = \frac{2}{\sqrt{\Omega_{\rm m} H_0^2}} \left[\sqrt{a + a_{\rm eq}} - \sqrt{a_{\rm eq}}\right]. \tag{2.95}$$

What is the conformal time at $z = 1100$?

**2.7** Consider a galaxy of physical size 5 kpc. What angle would this galaxy subtend if situated at redshift 0.1? Redshift 1? Do the calculation in a Euclidean universe, first matter-dominated and then for the fiducial concordance cosmology.

**2.8** How is the energy density of a gas of photons with a black-body spectrum related to the specific intensity of the radiation? That is, what is the relation between $\rho_\gamma$ and $I_\nu$ defined in Eq. (1.9)?

**2.9** **(a)** Compute the pressure of a relativistic species in equilibrium with temperature $T$. Show that $\mathcal{P} = \rho/3$ for both Fermi–Dirac and Bose–Einstein statistics.

**(b)** Suppose the distribution function depends only on $E/T$ as it does in equilibrium in the absence of a chemical potential. Find $d\mathcal{P}/dT$. A simple way to do this is to rewrite $df/dT$ in the integral as $-(E/T)df/dE$ and then integrate Eq. (2.64) by parts.

**2.10** Plot $d_L(z)$, $d_A(z)$, and $m - M$ as a function of redshift for a Euclidean, matter-dominated universe (this can be done analytically) and for the fiducial Euclidean concordance cosmology (for this you need to evaluate numerically a 1D integral). Neglect the $K$ correction for $m - M$. Compare with Fig. 2.6.

**2.11** Consider the entropy density, $s$, defined in Eq. (2.70). For a massless particle, you showed in Exercise 2.9 that $\mathcal{P} = \rho/3$, so $s = 4\rho/3T$. Express $s$ as a function of $T$ for both bosons and fermions (assumed massless) in equilibrium with zero chemical potential. Show that the entropy density for a massive particle in equilibrium ($T \ll m$; $\mu = 0$) is exponentially suppressed.

**2.12** Show that the number density of one generation of neutrinos and anti-neutrinos in the universe today is

$$n_{\nu_i} = \frac{3}{11} n_\gamma = 112\ {\rm cm}^{-3}.$$

For this calculation, you will also have to compute the photon number density; both $n_{\nu_i}$ and $n_\gamma$ can be expressed in terms of Riemann zeta functions (Eq. (C.30)). Using this result, verify Eq. (2.84).

**2.13** Consider the following two scenarios. Each has energy density equal to the critical density divided up between only two components: a cold, dark matter particle and a neutrino. The neutrino in each case has the standard abundance and temperature. The only difference between the two scenarios is in one the neutrino is massless while in the other it has a mass of 0.06 eV. Plot the energy density as a function of scale factor in each of these scenarios. Note that they should agree very early on (in each case there is only a relativistic neutrino early on) and very late. The only difference comes in the middle.

**2.14** Determine the epoch of dark energy-matter equality assuming that the dark energy is a cosmological constant.

# 3

# The fundamental equations of cosmology

Cosmology is, essentially, an application of general relativity coupled with statistical mechanics. The only relevant long-range force is gravity, which also provides the background spacetime within which matter moves, as we have seen in the last chapter. Since cosmology deals with the evolution of the entire universe, we are not interested in the fate of individual particles. Instead, we care about the collective, average behavior of matter, which is described by statistical mechanics. This is why essentially all results in cosmology can be derived from the combination of two equations: the Einstein equations on the gravity side, and the Boltzmann equations of statistical mechanics for matter and radiation.

These are formidable equations, and their application can quickly get technical. In this chapter, we will present the general form of the Einstein and Boltzmann equations, and describe their physical content. We will then apply them to the homogeneous universe, which, for the Einstein equations, allows us to derive the Friedmann equation (1.3). These results will also allow us to compute the expansion history and thermal history of the universe in this chapter and the next. Further, with the experience we gain in this chapter, there will be nothing particularly difficult about the subsequent chapters which deal with perturbations in the universe. So, becoming familiar with the framework laid out in this chapter will pay off greatly when going through the rest of the book.

## 3.1 Einstein equations

In the previous chapter, we have dealt with gravity only in terms of the metric, which gives us a notion of distances and straight lines (geodesics) in general spacetimes. These results were built on the principle of general covariance alone. We now turn to the second aspect of general relativity, which relates the metric to the constituents of the universe. This second part is contained in the Einstein equations, which relate the Einstein tensor describing the geometry to the energy-momentum tensor of matter.[1] This set of equations can be summarized as the following celebrated tensor equality (Fig. 3.1):

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{3.1}$$

[1] Relativists often refer to anything that appears on the right-hand side of the Einstein equations as "matter," even though this in general includes radiation and other constituents. We occasionally use this nomenclature too, so beware.

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00009-7
Copyright © 2021 Elsevier Inc. All rights reserved.

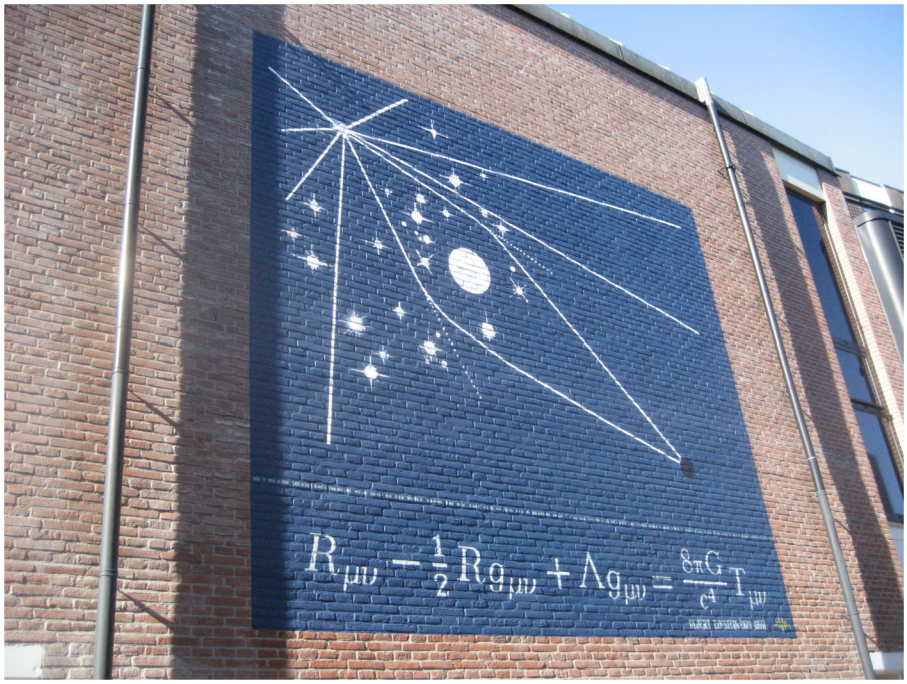



**FIGURE 3.1** The Einstein equations painted on the wall of a building in Leiden, the Netherlands. The drawing illustrates gravitational lensing, cf. Fig. 3.4. Painting by Jan-Willem Bruins (TegenBeeld); photograph by Vysotsky—Own work, CC BY-SA 4.0 (https://commons.wikimedia.org/w/index.php?curid=50130596).

Here $G_{\mu\nu}$ is the Einstein tensor defined through

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R. \tag{3.2}$$

$R_{\mu\nu}$ is the *Ricci tensor*, which depends only on the metric and its derivatives; $R$, the *Ricci scalar*, is the contraction of the Ricci tensor ($R \equiv g^{\mu\nu}R_{\mu\nu}$). Further, $\Lambda$ is the famous cosmological constant, $G$ is Newton's constant, and $T_{\mu\nu}$ is the energy-momentum tensor, whose expression in the background universe we have already encountered in Sect. 2.3. Thus, the left-hand side of Eq. (3.1) is a function of the metric, the right a function of the constituents of the universe: the Einstein equations relate the two.

The simplicity of Eq. (3.1) belies the rich physics encoded in the Einstein equations. They govern the evolution of the smooth universe as well as the growth of structure within it. On small scales, Newtonian gravity is included, as we will see in Sect. 3.3, as are black holes which we will not deal with in this book. We will later encounter a different purely general-relativistic effect contained in Eq. (3.1) though: gravitational waves.

The Ricci tensor is most conveniently expressed in terms of the Christoffel symbols [Eq. (2.21)],

$$R_{\mu\nu} = \Gamma^{\alpha}{}_{\mu\nu,\alpha} - \Gamma^{\alpha}{}_{\mu\alpha,\nu} + \Gamma^{\alpha}{}_{\beta\alpha}\Gamma^{\beta}{}_{\mu\nu} - \Gamma^{\alpha}{}_{\beta\nu}\Gamma^{\beta}{}_{\mu\alpha}. \tag{3.3}$$

Here commas denote derivatives with respect to $x$. So, for example, $\Gamma^{\alpha}{}_{\mu\nu,\alpha} \equiv \partial\Gamma^{\alpha}{}_{\mu\nu}/\partial x^{\alpha}$.

Let us now assume the FLRW form of the metric as an *ansatz* and derive the Einstein equations Eq. (3.1) for a metric of this form. The (curved) FLRW metric is in fact the most general form of metric that is spatially homogeneous, and thus is the appropriate ansatz for the homogeneous universe. While the FLRW metric in general has curved spatial slices, we assume a Euclidean universe throughout this derivation as it illustrates the essential features. We leave it as an exercise (Exercise 3.5) to derive the Einstein equations in a curved universe. Although Eq. (3.3) looks formidable, we have already done a big part of the work by computing the Christoffel symbol in an FLRW universe in Sect. 2.1.2.

Before embarking on the computation, let us pause to think about what we expect. We know that the Christoffel symbols are proportional to the first derivative of the metric with respect to the coordinates. Thus, from the structure of Eq. (3.3), the Ricci tensor involves $(i)$ terms that are proportional to the second derivatives of the metric, and $(ii)$ terms that involve the first derivatives squared. Due to the simplicity of the FLRW metric (for the Euclidean case we assume here), which involves only a single function of time $a(t)$, we immediately see that all components of $R_{\mu\nu}$ are either proportional to $(i)$ $\ddot{a}$ or $(ii)$ $\dot{a}^2$. We can say even more about the Ricci scalar $R$: as a scalar, it cannot depend on our choice of coordinates. But the absolute value of $a(t)$ can always be changed by multiplying the spatial coordinates by a constant; recall that we chose them such that $a(t_0) = 1$. So, $R$ cannot depend on $\ddot{a}$ but must involve $\ddot{a}/a$; similarly, it can only depend on $(\dot{a}/a)^2 = H^2$. Our computation then comes down to only computing the constants in front of each term. This type of symmetry consideration is very useful to check whether the results of a computation make physical sense. So let us proceed.

By working through the math, we find that there are only two sets of nonvanishing components of the Ricci tensor: one with $\mu = \nu = 0$ and the other with $\mu = \nu = i$. Consider

$$R_{00} = \Gamma^{\alpha}{}_{00,\alpha} - \Gamma^{\alpha}{}_{0\alpha,0} + \Gamma^{\alpha}{}_{\beta\alpha}\Gamma^{\beta}{}_{00} - \Gamma^{\alpha}{}_{\beta 0}\Gamma^{\beta}{}_{0\alpha}. \tag{3.4}$$

Recall that in our case the Christoffel symbol vanishes if its two lower indices are zero, so the first and third terms on the right vanish. Similarly, the indices $\alpha$ and $\beta$ in the second and fourth terms must be spatial. We are left with

$$R_{00} = -\Gamma^{i}{}_{0i,0} - \Gamma^{i}{}_{j0}\Gamma^{j}{}_{0i}. \tag{3.5}$$

Using Eq. (2.25) leads directly to

$$\begin{aligned} R_{00} &= -\delta_{ii}\frac{\partial}{\partial t}\left(\frac{\dot{a}}{a}\right) - \left(\frac{\dot{a}}{a}\right)^2 \delta_{ij}\delta_{ij} \\ &= -3\left[\frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2}\right] - 3\left(\frac{\dot{a}}{a}\right)^2 \\ &= -3\frac{\ddot{a}}{a}. \end{aligned} \tag{3.6}$$

The factors of 3 on the second line arise since $\delta_{ij}\delta_{ij} = \delta_{ii}$ signifies a sum over all three spatial indices, counting 1 for each. The space-space component is left as an exercise; it is

$$R_{ij} = \delta_{ij}[2\dot{a}^2 + a\ddot{a}]. \tag{3.7}$$

The next ingredient in the Einstein equations is the Ricci scalar, which we can now compute since

$$\begin{aligned} R &\equiv g^{\mu\nu}R_{\mu\nu} \\ &= -R_{00} + \frac{1}{a^2}R_{ii}. \end{aligned} \tag{3.8}$$

Again the sum over $i$ leads to a factor of 3, so

$$R = 6\left[\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2\right]. \tag{3.9}$$

As expected, it contains only $\ddot{a}/a$ and $(\dot{a}/a)^2$. Beyond cosmology, the Ricci scalar is useful to figure out whether a given metric describes a curved space or just Euclidean space written in strange coordinates (Exercise 3.1).[2]

We are now ready to write down the Einstein equations. Before that, let us do one final manipulation involving the cosmological constant. Nothing, of course, prevents us from moving the $\Lambda$ term to the right-hand side of Eq. (3.1), since it just involves the metric tensor. We can then formally define the cosmological-constant contribution to the stress-energy tensor:

$$T_{(\Lambda)}{}^{\mu}{}_{\nu} = -\frac{\Lambda}{8\pi G}\delta^{\mu}_{\nu} = \begin{pmatrix} -\rho_\Lambda & 0 & 0 & 0 \\ 0 & -\rho_\Lambda & 0 & 0 \\ 0 & 0 & -\rho_\Lambda & 0 \\ 0 & 0 & 0 & -\rho_\Lambda \end{pmatrix}, \quad \text{where} \quad \rho_\Lambda \equiv \frac{\Lambda}{8\pi G} \tag{3.10}$$

is the effective energy density of the cosmological constant. We see then that $\mathcal{P}_\Lambda = -\rho_\Lambda$, or in other words, $w_\Lambda = -1$, where we have used the energy-momentum tensor in Eq. (2.44)

[2] Note, however, that there are famous examples of curved spaces where $R$ happens to vanish, such as the Schwarzschild black hole solution where $R$ is zero away from the singularity.

to make this identification. We have thus shown that the equation of state of the cosmological constant is exactly $-1$. From the evolution of the energy density (Eq. (2.61)) we know that this had to be so: for any other $w_\Lambda$, the effective energy density of the cosmological constant would not be *constant*. The big advantage of including $\Lambda$ on the right-hand side of the Einstein equations is that it is then very easy to generalize all results to a non-$\Lambda$, dynamical form of dark energy.

Let us then proceed to derive the equation for the scale factor in a homogeneous universe; in fact, we need to consider only the time-time component of the Einstein equations:

$$R_{00} - \frac{1}{2} g_{00} R = 8\pi G T_{00}. \tag{3.11}$$

The terms on the left sum to $3\dot{a}^2/a^2$, and the component $T_{00}$ of the energy-momentum tensor is simply the energy density $\rho$. So we finally have

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho, \tag{3.12}$$

where recall that $\rho$ now includes $\rho_\Lambda$ (or, in general, $\rho_{\rm DE}$). This is termed the *first Friedmann equation* (in Exercise 3.4 you will derive the second Friedmann equation). In case this was your first time deriving an explicit form of the Einstein equations: congratulations! This is not an easy task in general.

To get this equation into a form closer to Eq. (1.3), recall that $(\dot{a}/a)^2$ is the square of the Hubble rate and that the critical density was defined as $\rho_{\rm cr} \equiv 3H_0^2/8\pi G$. So, dividing both sides by $H_0^2$ leads to

$$\frac{H^2(t)}{H_0^2} = \frac{\rho(t)}{\rho_{\rm cr}} = \sum_{s={\rm r,m},\nu,{\rm DE}} \Omega_s \left[a(t)\right]^{-3(1+w_s)}. \tag{3.13}$$

Here the energy density $\rho$ counts the energy density from all species: matter, radiation, neutrinos, and dark energy. In the second equality, we have used Eq. (2.71) and assumed a constant equation of state for all components (this is not quite correct at least for neutrinos, but the generalization is simple). In our derivation, we have assumed that the universe is Euclidean. Hence, Eq. (3.13) does not contain a term corresponding to the curvature of the universe. This is simple to add following Eq. (1.3). Defining $\Omega_{\rm K} \equiv 1 - \Omega_0 \equiv 1 - \sum_s \Omega_s$, we have

$$\frac{H^2(t)}{H_0^2} = \sum_{s={\rm r,m},\nu,{\rm DE}} \Omega_s \left[a(t)\right]^{-3(1+w_s)} + \Omega_{\rm K} \left[a(t)\right]^{-2}. \tag{3.14}$$

This equation is all we need to calculate the evolution of the homogeneous universe.

In Sect. 3.3.1, we will go beyond the background universe to include perturbations. We will then encounter a whole host of additional gravitational physics, including the generalization of Newton's theory of gravity to an expanding universe.

## 3.2 Boltzmann equation

After having treated gravity in the homogeneous universe, let us now turn to the equations governing matter and radiation. In cosmology, we are not interested in the fate of individual particles, but in their behavior in a *statistical* sense. Hence let us consider a collection of particles occupying some region of space, as we did in Sect. 2.3. In classical physics, these particles are completely described by the set $\{\boldsymbol{x}_i, \boldsymbol{p}_i\}$ of their positions $\boldsymbol{x}_i$ and momenta $\boldsymbol{p}_i$. We can then define the distribution function, as in Sect. 2.3, by relating it to the number of particles in a small phase-space element around $(\boldsymbol{x}, \boldsymbol{p})$:

$$N(\boldsymbol{x}, \boldsymbol{p}, t) = f(\boldsymbol{x}, \boldsymbol{p}, t)(\Delta x)^3 \frac{(\Delta p)^3}{(2\pi)^3}. \tag{3.15}$$

In the limit of a large number of particles within the volume element considered, $f(\boldsymbol{x}, \boldsymbol{p}, t)$ approaches a continuous function describing the state of the collection of particles, and we no longer need to keep track of individual particles. We already saw that the appropriate integration measure (in natural units) is given by $d^3x d^3p/(2\pi)^3$. Note that we do not need to include the energy as a separate variable, since, at any point in phase space, $E$ is completely determined by $(\boldsymbol{x}, \boldsymbol{p})$.

Now we would like to derive an equation governing this distribution function. This equation should uniquely follow from the equations of motion obeyed by the individual particles. Let us begin by neglecting any particle–particle interactions. Then, the only forces acting on the particles are long-range forces, which we can describe through a force field (more precisely, acceleration field) $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{p}, t)$. This could for example be gravity, in which case $\boldsymbol{a} = -\nabla\Psi(\boldsymbol{x}, t)$, where the gravitational potential $\Psi$ (defined in Eq. (3.49) below) is independent of the particle momenta, or it could be the Lorentz force due to electromagnetic fields. Then, using the definition of the momentum $\boldsymbol{p}$, the equations of motion for nonrelativistic particles are

$$\dot{\boldsymbol{x}} = \frac{\boldsymbol{p}}{m}; \quad \dot{\boldsymbol{p}} = m\,\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{p}, t). \tag{3.16}$$

The number of particles is conserved, which we can formalize by stating that the total time derivative of $f$ vanishes,

$$\frac{df(\boldsymbol{x}, \boldsymbol{p}, t)}{dt} = 0 \quad \text{where} \quad \frac{d}{dt} = \frac{\partial}{\partial t} + \dot{\boldsymbol{x}} \cdot \nabla_x + \dot{\boldsymbol{p}} \cdot \nabla_p \tag{3.17}$$

is the total (rather than partial) time derivative, and $\nabla_x$, $\nabla_p$ denote the gradient with respect to the arguments $\boldsymbol{x}$ and $\boldsymbol{p}$, respectively. Inserting the equations of motion, this becomes

$$\begin{aligned} \frac{\partial f(\boldsymbol{x}, \boldsymbol{p}, t)}{\partial t} &= -\dot{\boldsymbol{x}} \cdot \nabla_x f(\boldsymbol{x}, \boldsymbol{p}, t) - \dot{\boldsymbol{p}} \cdot \nabla_p f(\boldsymbol{x}, \boldsymbol{p}, t) \\ &= -\frac{\boldsymbol{p}}{m} \cdot \nabla_x f(\boldsymbol{x}, \boldsymbol{p}, t) - m\,\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{p}, t) \cdot \nabla_p f(\boldsymbol{x}, \boldsymbol{p}, t). \end{aligned} \tag{3.18}$$

That is, the rate of change $\partial f/\partial t$ of the distribution function is determined by how many particles move in and out of the phase-space volume element considered, or equivalently that the phase-space volume occupied by a collection of particles is conserved (Fig. 3.2). The catch is that these particles themselves are moving through phase space in complicated ways in general. This catch makes the problem more difficult than it seems from Eq. (3.17).

Further, if particle–particle interactions are relevant, then this equation must be modified to include a source term, a *collision term*, on the right-hand side, which describes how particles are moved from one phase-space element to another (typically at the same position $\boldsymbol{x}$):

$$\frac{df}{dt} = C[f]. \tag{3.19}$$

We will turn to deriving such collision terms in Sect. 3.2.3.

The payoff of all this work is that the distribution function allows us to derive all macroscopic properties of the collection of particles, such as density and pressure, as we have already seen in Sect. 2.3. Of particular importance is the energy-momentum tensor $T_{\mu\nu}$, since it appears on the right-hand side of the Einstein equations Eq. (3.1). The relativistic expression for the energy-momentum tensor given a distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$ is

$$T^{\mu}{}_{\nu}(\boldsymbol{x}, t) = \frac{g}{\sqrt{-\det[g_{\alpha\beta}]}} \int \frac{dP_1 dP_2 dP_3}{(2\pi)^3} \frac{P^{\mu} P_{\nu}}{P^0} f(\boldsymbol{x}, \boldsymbol{p}, t), \tag{3.20}$$

where the degeneracy factor $g$ counts how many different particle states are in fact described by the distribution function $f$. $P^{\mu} = dx^{\mu}/d\lambda$ is the *comoving momentum* defined in Eq. (2.26) and $P_{\mu} = g_{\mu\nu} P^{\nu}$, while the *physical momentum* $\boldsymbol{p}$ is related to $P^i$ via Eq. (2.32). We can raise and lower indices on this tensor by acting on $P^{\mu} P_{\nu}$ with the metric. The derivation of Eq. (3.20) is quite subtle (Ma and Bertschinger, 1995), so let us just briefly go through the different factors in this equation. The energy-momentum tensor essentially gives the current density of the 4-momentum carried by the particles with distribution function $f$. The momentum integral over $f$, which gives the particle number density, weighted by $P_{\nu}$ yields the 4-momentum density. In order to obtain the current, we have to multiply by the velocity $P^{\mu}/P^0$, just like in the case of a charge current $\boldsymbol{j} = n\boldsymbol{v}$. Then, the prefactor involving the determinant of the metric is essentially a geometric factor which is necessary in order to ensure that $T^{\mu}{}_{\nu}$ obeys the correct conservation law: $\nabla_{\mu} T^{\mu}{}_{\nu} = 0$. In Exercise 3.7, you will show that in the homogeneous universe, Eq. (3.20) leads to the relations for the energy density and pressure we have introduced in Sect. 2.3.

### 3.2.1 Boltzmann equation for particles in a harmonic potential

Let us begin our journey with the Boltzmann equation with the case of nonrelativistic particles governed by a simple $x^2$ potential in one dimension. This Boltzmann equation exhibits all essential features of the full general-relativistic versions of the Boltzmann equation we will encounter in the next section, but the algebra is much less cumbersome. So

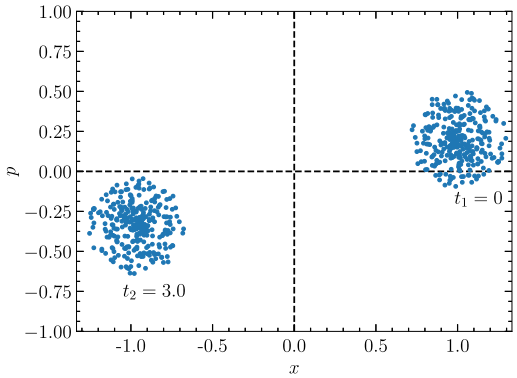



**FIGURE 3.2** Phase space for a set of collisionless particles in a harmonic potential. The initial distribution at $t_1$ moves in phase space to reach a different position at time $t_2$. The phase-space volume occupied by the particles is conserved throughout the evolution. In case of the harmonic potential, even the shape of the phase-space volume remains the same, while in general it gets distorted in the course of time evolution.

here the physics will be quite transparent. It will be useful to keep this example in mind when the algebra threatens to obscure the physics in the following chapters.

Consider free particles living in a one-dimensional harmonic potential well. Their energy then is simply

$$E = \frac{p^2}{2m} + \frac{1}{2}kx^2, \tag{3.21}$$

where $k$ is the spring constant. The distribution function is now a function of three scalar arguments $f = f(x, p, t)$. Fig. 3.2 illustrates the movement through phase space of a distribution of such particles (throughout, we consider the collisionless case $C[f] = 0$). The full time derivative $df/dt$ vanishes since the number of particles in the bunch at $t_1$ equals that at $t_2$. What changes over time is the location of the particles in phase space themselves. Alternatively, we can think of $x$ and $p$ as independent variables (not dependent on $t$) and take partial derivatives of $f$ with respect to $t$, $x$, and $p$. All of these partial derivatives are nonzero, but the appropriate weighted sum of the three vanishes [Eq. (3.17)].

To determine the coefficients $\dot{x}$ and $\dot{p}$ in Eq. (3.17), we must use the equations of motion, i.e. the one-dimensional version of Eq. (3.16). Via Newton's force law, we have

$$\dot{x} = \frac{p}{m} \quad \text{and} \quad \dot{p} = -kx. \tag{3.22}$$

When generalizing to the relativistic case, these familiar equations will be replaced by the geodesic equation we have derived in Sect. 2.1.2. The collisionless Boltzmann equation for the present case is then

$$\frac{\partial f}{\partial t} + \frac{p}{m}\frac{\partial f}{\partial x} - kx\frac{\partial f}{\partial p} = 0. \tag{3.23}$$

The second term here governs how rapidly the particle moves in real space; the coefficient in front is just the velocity, $v = p/m$. The last term governs how quickly particles lose or gain momentum.

In order to solve the Boltzmann equation, which is a partial differential equation in the three variables $(x, p, t)$, we need to know the initial conditions on the distribution function. Even without these, though, the Boltzmann equation offers some useful insights. Consider a static distribution, defined by $\partial f/\partial t = 0$ (as opposed to $df/dt = 0$, which always holds). Such a distribution, also called an *equilibrium distribution*, means that the number of particles with a given momentum $p$ stays the same in a statistical sense at each point in space $x$. Of course, this does not mean that the particles themselves don't move. A general solution for the equilibrium distribution is

$$f(p, x) = f_{\rm EQ}(E[p, x]); \tag{3.24}$$

that is, $f$ is a function only of energy $E$. To see that this is indeed a solution, consider

$$\begin{aligned} \frac{p}{m}\frac{\partial f(E)}{\partial x} - kx\frac{\partial f(E)}{\partial p} &= \frac{df}{dE}\left[\frac{p}{m}\frac{\partial E}{\partial x} - kx\frac{\partial E}{\partial p}\right] \\ &= 0, \end{aligned} \tag{3.25}$$

where the second equality follows from Eq. (3.21). So any function of the energy alone is an equilibrium distribution. In the absence of collisions, as we have assumed here, which equilibrium distribution is the correct one depends entirely on the initial conditions. However, if there are interactions, then an equilibrium distribution also needs to make the collision term vanish: $C[f_{\rm EQ}] = 0$. This will in general drive $f$ to one of the familiar equilibrium distributions that we introduced in Sect. 2.3.

### 3.2.2 Boltzmann equation in an expanding universe

So far, we have studied the Boltzmann equation in the Minkowski-space context (cf. Eq. (3.16)), as appropriate for lab experiments on Earth. Let us now derive the generalization to an expanding spacetime. As we know from Sect. 2.1.2, the equations of motion Eq. (3.16) get generalized to the geodesic equation, and the three-momentum $\boldsymbol{p}$ is correspondingly promoted to a four-vector

$$P^\mu \equiv \frac{dx^\mu}{d\lambda} \tag{3.26}$$

where $\lambda$ again parametrizes the particle's path, as in Eq. (2.20) (and again we will not need to specify $\lambda$ explicitly). However, the distribution function for a given collection of particles remains a function defined on a six-dimensional phase space: first, we keep track of time separately as before, and second, the four-momentum of each particle obeys the *mass-shell constraint*

$$P^2 \equiv g_{\mu\nu}P^\mu P^\nu = -m^2, \tag{3.27}$$

where $m$ is the particle rest mass (which could be zero, e.g. for photons). Defining the norm of the three-momentum $p$, by generalizing Eq. (2.32) to

$$p^2 \equiv g_{ij} P^i P^j, \tag{3.28}$$

Eq. (3.27) becomes, for the FLRW metric,

$$E^2 \equiv (P^0)^2 = p^2 + m^2. \tag{3.29}$$

Thus, we have eliminated $P^0$ in favor of $\boldsymbol{p}$, and we can write a relativistic Boltzmann equation for a distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$ as before. It is convenient to separate the dependence on $\boldsymbol{p}$ into a dependence on its magnitude $p \equiv \sqrt{p^2}$ and its unit vector $\hat{p}^i = \hat{p}_i$, which satisfies $\delta_{ij} \hat{p}^i \hat{p}^j = 1$ by definition. We expect that $\hat{p}^i$ is proportional to the comoving momentum $P^i$; call the proportionality constant $C$:

$$P^i \equiv C \hat{p}^i. \tag{3.30}$$

To determine the coefficient $C$, we can use Eq. (3.28):

$$\begin{aligned} p^2 &= g_{ij} \hat{p}^i \hat{p}^j C^2 \\ &= a^2 \delta_{ij} \hat{p}^i \hat{p}^j C^2 \\ &= a^2 C^2. \end{aligned} \tag{3.31}$$

Eq. (3.31) tells us that $C = p/a$, so whenever we encounter $P^i$, we can always eliminate it in favor of $p$, $\hat{p}^i$ via

$$P^i = \frac{p}{a} \hat{p}^i. \tag{3.32}$$

We can now write Eq. (3.17) as

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x^i} \cdot \frac{dx^i}{dt} + \frac{\partial f}{\partial p} \frac{dp}{dt} + \frac{\partial f}{\partial \hat{p}^i} \cdot \frac{d\hat{p}^i}{dt}. \tag{3.33}$$

Again, in this section we are attempting to derive the Boltzmann equation only for the smooth, expanding universe. As we have seen in Sect. 2.1.2, the direction of the particle momentum does not change in an expanding universe (even though its magnitude does). Thus, we can immediately drop the last term, $\propto d\hat{p}^i/dt$, in Eq. (3.33).

Next let us reexpress the second term on the right-hand side, $\propto dx^i/dt$, by recalling that $P^i \equiv dx^i/d\lambda$ and $P^0 \equiv dt/d\lambda$. Therefore,

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{dx^i}{d\lambda} \frac{d\lambda}{dt} \\ &= \frac{P^i}{P^0} = \frac{p}{E} \frac{\hat{p}^i}{a}, \end{aligned} \tag{3.34}$$

where we have used Eq. (3.29) and Eq. (3.32). Next up is an equation for $dp/dt$. For this, let us recall that the time component of the geodesic equation (2.20) can be written as

$$\frac{dP^0}{d\lambda} = -\Gamma^0{}_{\alpha\beta} P^\alpha P^\beta. \tag{3.35}$$

We can rewrite the derivative with respect to $\lambda$ as a derivative with respect to time multiplied by $dt/d\lambda = P^0$. Using the Christoffel symbols for the FLRW metric [Eq. (2.21)], we obtain

$$P^0 \frac{dP^0}{dt} = -\Gamma^0{}_{ij} P^i P^j. \tag{3.36}$$

The mass-shell relation Eq. (3.29), written in the form $P^0 dP^0/dt = (1/2)d(E^2)/dt$, and Eq. (2.24) then lead to

$$p\frac{dp}{dt} = -Hp^2 \quad \Rightarrow \quad \frac{dp}{dt} = -Hp. \tag{3.37}$$

This equation says that the physical momentum of any particle decays as $1/a$ in an unperturbed expanding universe, a fact we already know from the previous chapter. We finally obtain the Boltzmann equation in the homogeneous expanding universe:

$$\frac{\partial f}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - Hp\frac{\partial f}{\partial p} = C[f]. \tag{3.38}$$

One might wonder why we have kept the contribution from $\partial f/\partial x^i$ here, as it is unnecessary given the assumption of a homogeneous universe. The answer is simply that keeping this term was easy enough, and we will need its generalization later when dealing with perturbations.

Eq. (3.38) is valid for *all particles*. However, we will frequently encounter two limits of these Boltzmann equations: in the *relativistic* limit, when $p \gg m$, we have $E \simeq p$ and thus obtain

$$\frac{\partial f}{\partial t} + \frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - Hp\frac{\partial f}{\partial p} = C[f]. \tag{3.39}$$

This applies to photons as well as neutrinos while they are relativistic. In the opposite *non-relativistic* limit, $p \ll m$ so that $E \simeq m$. Eq. (3.38) becomes

$$\frac{\partial f}{\partial t} + \frac{p}{m}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - Hp\frac{\partial f}{\partial p} = C[f]. \tag{3.40}$$

Note in particular that the coefficient in front of the second term is now small, suppressed by the speed $|\boldsymbol{v}| = p/m$.

Now let us derive how the Boltzmann equation helps us to calculate the evolution of the number density of the species under consideration. The number density $n(\boldsymbol{x}, t)$ is simply

given by the integral of $f(\boldsymbol{x}, \boldsymbol{p}, t)$ over all momenta. So let us integrate Eq. (3.38) over $\boldsymbol{p}$, using the fact that in the homogeneous universe, $\partial f/\partial x^i = 0$:

$$\int \frac{d^3p}{(2\pi)^3}\frac{\partial f}{\partial t} - H\int \frac{d^3p}{(2\pi)^3}p\frac{\partial f}{\partial p} = \int \frac{d^3p}{(2\pi)^3}C[f]. \tag{3.41}$$

The second term on the left-hand side can be dealt with by integrating by parts:

$$\int \frac{d^2\hat{p}}{(2\pi)^3}\int_0^\infty p^2dp\, p\frac{\partial f}{\partial p} = -3\int \frac{d^2\hat{p}}{(2\pi)^3}\int_0^\infty p^2dpf(\boldsymbol{p}), \tag{3.42}$$

where we have used that, for any regular distribution function, $p^3 f(\boldsymbol{p})$ vanishes at $p = 0$ as well as at infinity. Eq. (3.41) becomes

$$\frac{dn(t)}{dt} + 3Hn(t) = \int \frac{d^3p}{(2\pi)^3}C[f]. \tag{3.43}$$

In the absence of collisions, the particle number decays as $a^{-3}$, as we already knew: as the comoving grid expands, the volume of a region containing a fixed number of particles grows as $a^3$. Therefore, the physical number density of these particles falls off as $a^{-3}$. However, collisions can change that behavior if the integral over momentum of the collision term does not vanish. We will consider collisions next.

### 3.2.3 Collision terms

The effect of direct particle interactions is, in the Boltzmann realm, referred to as "collisions." Collisions include scattering as well as pair creation, annihilation, and particle decay. A common type of process is a reaction where particles of type 1 and 2 interact to form particles of type 3 and 4:

$$(1)_{\boldsymbol{p}} + (2)_{\boldsymbol{q}} \longleftrightarrow (3)_{\boldsymbol{p}'} + (4)_{\boldsymbol{q}'}, \tag{3.44}$$

where the subscripts indicate momenta. Note that this includes scattering of electrons and photons for example, if we choose $(1) = (3) = (e^-)$ and $(2) = (4) = (\gamma)$; or annihilation, if we choose $(1) = (e^-)$, $(2) = (e^+)$ and $(3) = (4) = (\gamma)$. Moreover, all microscopic physical processes conserve momentum and energy:

$$\boldsymbol{p} + \boldsymbol{q} = \boldsymbol{p}' + \boldsymbol{q}'; \quad E_1(\boldsymbol{p}) + E_2(\boldsymbol{q}) = E_3(\boldsymbol{p}') + E_4(\boldsymbol{q}'), \tag{3.45}$$

where $E_s(p) = \sqrt{p^2 + m_s^2}$ denotes the energy-momentum relation for particle $s$ [Eq. (3.29)]. Each type of particle has its respective distribution function $f_s(\boldsymbol{x}, \boldsymbol{p}, t)$, $s = 1, 2, 3, 4$. Often in cosmology, different states (e.g. spin) have the same distribution function. So, instead of following them with separate functions, we will assign appropriate statistical weights $g_s$.

How does the reaction Eq. (3.44) affect the evolution of the distribution functions $f_s$ of the particles involved? First, we are dealing with a local interaction in space and time, so all

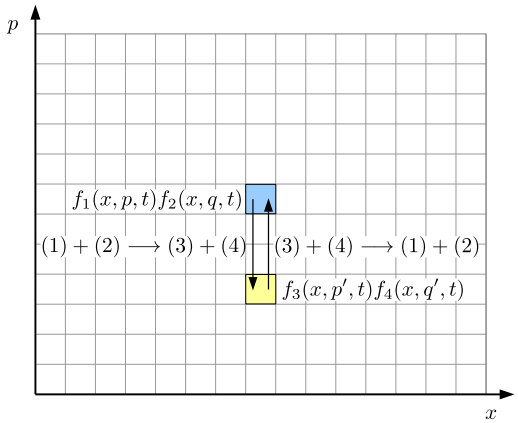



**FIGURE 3.3** Illustration of the effect of collisions on the phase-space distribution function for particle (1), i.e. the collision term $C[f_1(p,x,t)]$, in a 1D setting. Consider the blue (dark shaded) cell at $x$, $p$. The forward reaction in Eq. (3.44) removes particles from $f_1(x,p,t)$ in proportion to the product $f_1 f_2$, and adds them to the distribution function for particle (3) and (4) in the lower light-shaded cell. The reverse reaction on the other hand adds to $f_1(x,p,t)$, in proportion to the abundance of (3) and (4) particles in the lower light-shaded cell. All collisions happen locally, i.e. at a fixed position $x$. We have assumed that the particular momenta shown are kinematically allowed, and not included the factors for stimulated emission or Pauli blocking here for simplicity.

the distribution functions are evaluated at $(\boldsymbol{x},t)$, and we only need to determine the momentum arguments. For $f_1(\boldsymbol{x},\boldsymbol{p},t)$, for example, Eq. (3.44) means that we have to subtract the particles of type 1 that get scattered away from momentum $\boldsymbol{p}$ by the forward reaction, and add the particles of type 1 that get scattered to momentum $\boldsymbol{p}$ by the reverse reaction (Fig. 3.3). Therefore we must sum over all other momenta $(\boldsymbol{q},\boldsymbol{q}',\boldsymbol{p}')$ which affect $f_1(\boldsymbol{p})$. Schematically, then, the collision term is

$$C[f_1(\boldsymbol{p})] = \sum_{\boldsymbol{q},\boldsymbol{q}',\boldsymbol{p}'}^{\boldsymbol{p}+\boldsymbol{q}=\boldsymbol{p}'+\boldsymbol{q}'} \delta_{\rm D}^{(1)}(E_1(p)+E_2(q)-E_3(p')-E_4(q'))\,|\mathcal{M}|^2 \times \{f_3(\boldsymbol{p}')f_4(\boldsymbol{q}')-f_1(\boldsymbol{p})f_2(\boldsymbol{q})\}, \tag{3.46}$$

where the Dirac delta function enforces energy conservation. Here, the scattering amplitude squared $|\mathcal{M}|^2$ depends on the microphysical details of the interaction and can be computed using Feynman diagrams. Since we are dealing with a 2-particle interaction, the scattering rate depends on the product of distribution functions $f_1 f_2$ (forward reaction) and $f_3 f_4$ (reverse reaction). Moreover, the amplitude for forward and reverse reactions is the same. Here and in the following, we no longer write the arguments $\boldsymbol{x}$, $t$ of the distribution functions, since they are all evaluated at the same $\boldsymbol{x}$ and $t$.

There is one ingredient we have neglected in Eq. (3.46): quantum effects such as stimulated emission (or Bose enhancement) and the Pauli exclusion principle (or Pauli blocking). These increase or suppress the reaction rate depending on the occupation of the *final* state. Including them amounts to adding factors of $(1 \pm f_3)(1 \pm f_4)$ to the forward reac-

tion, and $(1 \pm f_1)(1 \pm f_2)$ to the reverse reaction. In each case, a plus sign applies when the corresponding particle is a boson, and a minus sign when it is a fermion. Pauli blocking is particularly obvious: if the state of fermion 1 corresponding to momentum $\boldsymbol{p}$ is already occupied, the factor $(1 - f_1(\boldsymbol{p}))$ is zero, and the inverse reaction with this final state does not happen, as required. Conversely, if particle 1 is a boson, the corresponding reaction rate is enhanced, as bosons like to occupy the same state.

Finally, we can put in the appropriate factors in Eq. (3.46) to properly perform the sums over phase space. First, as in Fig. 2.4, the volume element in phase space is $d^3p/(2\pi)^3$ [really $d^3p/(2\pi\hbar)^3$]. Second, in the relativistic setting the phase-space integrals are four-dimensional, over the three components of momentum and the energy. However, the energy is constrained by the on-shell condition Eq. (3.29), which requires $E_s = (p^2 + m_s^2)^{1/2}$. We can thus perform the energy integral as follows:

$$\int d^3p \int_0^\infty dE\, \delta_{\rm D}^{(1)}(E^2 - p^2 - m^2) = \int d^3p \int_0^\infty dE \frac{\delta_{\rm D}^{(1)}\left(E - \sqrt{p^2+m^2}\right)}{2E}. \tag{3.47}$$

Performing the integral over $E$ with the delta function yields a factor of $1/2E$. To summarize, the infinitesimal phase-space volume element to integrate over for each particle $i$ is $d^3p_i/[(2\pi)^3 2E_i(p_i)]$. With this, the collision term becomes

$$\begin{aligned} C[f_1(\boldsymbol{p})] = \frac{1}{2E_1(p)} \int \frac{d^3q}{(2\pi)^3 2E_2(q)} \int \frac{d^3p'}{(2\pi)^3 2E_3(p')} \int \frac{d^3q'}{(2\pi)^3 2E_4(q')} |\mathcal{M}|^2 \\ \times (2\pi)^4 \delta_{\rm D}^{(3)}[\boldsymbol{p} + \boldsymbol{q} - \boldsymbol{p}' - \boldsymbol{q}']\, \delta_{\rm D}^{(1)}[E_1(p) + E_2(q) - E_3(p') - E_4(q')] \\ \times \Big\{ f_3(\boldsymbol{p}') f_4(\boldsymbol{q}') \left[1 \pm f_1(\boldsymbol{p})\right]\left[1 \pm f_2(\boldsymbol{q})\right] \\ - f_1(\boldsymbol{p}) f_2(\boldsymbol{q}) \left[1 \pm f_3(\boldsymbol{p}')\right]\left[1 \pm f_4(\boldsymbol{q}')\right] \Big\}. \end{aligned} \tag{3.48}$$

Again, the delta functions enforce energy and momentum conservation. This result holds in general for any 2-particle interaction of the type in Eq. (3.44), where several of the particles 1,2,3,4 could be of the same species. All the microphysical details of this interaction are encoded in the amplitude squared $|\mathcal{M}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{p}', \boldsymbol{q}')|^2$, which in general depends on the momenta of the particles involved. Eq. (3.48) can be straightforwardly generalized to processes involving fewer particles, such as particle decay, by assembling the collision term out of the amplitude, appropriate products of distribution functions, and integrals over momenta. We will encounter concrete examples in the next two chapters.

## 3.3 Beyond the homogeneous universe

So far, we have derived the Einstein and Boltzmann equations in the homogeneous universe. This is already sufficient to calculate the thermal history of the universe which is the topic of Ch. 4, including the production of dark matter, Big Bang Nucleosynthesis, and the formation of the first atoms. We will now go beyond the smooth universe and consider the

presence of inhomogeneities, which will occupy us for most of the rest of the book. Readers should feel free to jump ahead to Ch. 4, and come back to this section later.

Due to the simplicity of the smooth universe, we were able to get away without any approximations and derive the exact Einstein and Boltzmann equations. The universe with structure, however, is a far more complicated (and richer) case than the homogeneous universe. For this reason, we will have to rely on some approximations, the most important being that the deviations of the spacetime from the FLRW form are small. Fortunately, this approximation is very accurate in the realm of cosmology, as we will see.

### 3.3.1 Perturbed spacetime

To begin, we must specify the form of the metric, accounting for perturbations around the smooth universe described by Eq. (2.12). Whereas the smooth universe is characterized by a single function, $a(t)$, which depends only on time and not on space, the perturbed universe requires two more functions, $\Psi$ and $\Phi$, both of which depend on space and time. In terms of these, the metric can be written as

$$\begin{aligned} g_{00}(\boldsymbol{x},t) &= -1-2\Psi(\boldsymbol{x},t), \\ g_{0i}(\boldsymbol{x},t) &= 0, \\ g_{ij}(\boldsymbol{x},t) &= a^2(t)\delta_{ij}\left[1+2\Phi(\boldsymbol{x},t)\right]. \end{aligned} \tag{3.49}$$

In the absence of $\Psi$ and $\Phi$, Eq. (3.49) is simply the FLRW metric of the zeroth-order homogeneous, Euclidean cosmology. Conversely, in the absence of expansion ($a(t)=1$) this metric describes a weak gravitational field. The perturbations to the metric are $\Psi$, which corresponds to the Newtonian potential and governs the motion of slow-moving (nonrelativistic) bodies; and $\Phi$, the perturbation to the spatial curvature which, from Eq. (3.49), can also be interpreted as a local perturbation to the scale factor: $a(t) \to a(\boldsymbol{x},t) = a(t)\sqrt{1+2\Phi(\boldsymbol{x},t)}$. In general, there is a tight relation between $\Phi$ and $\Psi$, as we will see in later chapters.

The typical magnitude of metric perturbations $\Psi$, $\Phi$ in our universe is less than $10^{-4}$. For this reason, it is an excellent approximation to work *at linear order* in these quantities. This means that we neglect all terms that are quadratic or of higher order in them. We will work under this approximation, which greatly simplifies the calculations, throughout the entire book.

There are two technical points about the metric in Eq. (3.49) that you do not need to worry about for most of this book, but which nonetheless are important to be aware of. We will cover these issues in Ch. 6, when we study gravity in the inhomogeneous universe in more detail. First, one can break up perturbations into those behaving as scalars, vectors, and tensors under a transformation from one 3D spatial coordinate system to another. Eq. (3.49) contains only scalar perturbations. On the other hand, tensor perturbations correspond to gravitational waves, which we know to exist. To take these into account, $g_{\mu\nu}$ requires other functions besides $\Psi$ and $\Phi$. For now we focus solely on the scalar perturbations; these are by far the most important ones for the origin and evolution of structure in the universe.

The other feature of Eq. (3.49) worth noting is that its form corresponds to a particular choice of coordinates, or *gauge*. The simplest way to understand this gauge freedom is to think back to electricity and magnetism. There, the vector potential $A_\mu$ and its derivatives contain all possible information about the electric and magnetic fields. Since the physical $\boldsymbol{E}$ and $\boldsymbol{B}$ fields remain unchanged if the derivative of a scalar field, $\partial_\mu\varphi$, is added to $A_\mu$, there is some residual freedom in choosing the potential (for example, one often chooses $A_0 = 0$ or $\partial_\mu A^\mu = 0$). In our case of perturbations to the metric, a similar freedom exists. Even if only scalar perturbations are considered, there is still considerable freedom in the variables one can choose to describe the fluctuations. While any physical result must be independent of the gauge choice, it is possible to use a gauge that looks quite different from Eq. (3.49) and still describes the same physics. For the record, the gauge in Eq. (3.49) is called the *conformal Newtonian* gauge.

The first computation we did when we studied the FLRW metric Eq. (2.12) was to compute the Christoffel symbols [Eqs. (2.24–2.25)]. Now we need to do the same for the first-order terms, those that are linear in $\Phi$, $\Psi$. First let us consider ${\Gamma^0}_{\mu\nu}$, which can be written in terms of the metric as

$$ {\Gamma^0}_{\mu\nu} = \frac{1}{2} g^{0\alpha}[g_{\alpha\mu,\nu} + g_{\alpha\nu,\mu} - g_{\mu\nu,\alpha}] \tag{3.50} $$

where again $_{,\alpha}$ means the derivative with respect to $x^\alpha$. The only nonzero component of $g^{0\alpha}$ is the time component, which is the inverse of $g_{00} = -1 - 2\Psi$. So, to first order in the perturbations, $g^{00} = -1 + 2\Psi$, and

$$ {\Gamma^0}_{\mu\nu} = \frac{-1+2\Psi}{2}[g_{0\mu,\nu} + g_{0\nu,\mu} - g_{\mu\nu,0}]. \tag{3.51} $$

Take each component in turn: first the one with $\mu = \nu = 0$. Each of the terms in square brackets is identical, so the brackets give $g_{00,0} = -2\dot{\Psi}$. Since we are interested only in first-order terms, the factor of $2\Psi$ out in front can be dropped and we are left with

$$ {\Gamma^0}_{00} = \dot{\Psi}. \tag{3.52} $$

The next possibility is that one of the indices $\mu$ or $\nu$ is spatial and the other time. It doesn't matter which one is which, since the Christoffel symbol is symmetric in its lower indices. In this case, only one of the terms in brackets in Eq. (3.51) is nonzero, $g_{00,i} = -2\Psi_{,i}$. Once again, since this is of first order, we can drop the factor of $2\Psi$ in front, leading to

$$ {\Gamma^0}_{0i} = {\Gamma^0}_{i0} = \Psi_{,i}. \tag{3.53} $$

Finally, if both lower indices in Eq. (3.51) are spatial, the first two terms in brackets vanish since $g_{0i} = 0$ and only the last term survives, leaving

$$ {\Gamma^0}_{ij} = \frac{1-2\Psi}{2}\frac{\partial}{\partial t}[\delta_{ij}a^2(1+2\Phi)]. \tag{3.54} $$

There is a zeroth-order term here, the one we computed in Eq. (2.24), and three first-order terms:

$$\Gamma^0{}_{ij} = \delta_{ij} a^2 [H + 2H(\Phi - \Psi) + \dot{\Phi}] \tag{3.55}$$

with $H = \dot{a}/a$.

Computing the Christoffel symbols $\Gamma^i{}_{\mu\nu}$, will be left as an exercise. They are

$$\begin{aligned}
\Gamma^i{}_{00} &= \frac{1}{a^2}\Psi_{,i}, \\
\Gamma^i{}_{j0} &= \Gamma^i{}_{0j} = \delta_{ij}(H + \dot{\Phi}), \\
\Gamma^i{}_{jk} &= [\delta_{ij}\partial_k + \delta_{ik}\partial_j - \delta_{jk}\partial_i]\Phi.
\end{aligned} \tag{3.56}$$

Note that, at zeroth order, the only nonvanishing component is $\Gamma^i{}_{j0}$, in agreement with Eq. (2.25). By convention, both $\delta_{ij}$ and spatial derivatives $\partial_k$ live in Euclidean space, so we can freely interchange their upper and lower indices. This is standard convention in the cosmology literature when working with a Euclidean background universe.

You can already guess what we have to do next in order to obtain the desired Einstein equations for gravity: compute the Ricci tensor Eq. (3.3) and Ricci scalar for the perturbed metric to obtain the left-hand side; and compute the perturbed energy-momentum tensor for the right-hand side. Indeed, nothing stops us from doing those calculations right away. However, we will defer this to Ch. 6 for two reasons: first, it is useful to think a little more carefully about how to parametrize the tensors $G_{\mu\nu}$ and $T_{\mu\nu}$ in a perturbed universe; second, we can already get quite far with the Boltzmann equation in an expanding background alone. So let us continue with the latter.

### 3.3.2 The geodesic equation

In order to derive the Boltzmann equation, we need to know how particles move within the perturbed spacetime. Again, this is determined by the geodesic equation which we considered in Sect. 2.1.2, and which we now extend to include the spacetime perturbations $\Phi, \Psi$. In particular, our goal is to calculate $dx^i/dt$, $dp/dt$, and $d\hat{p}^i/dt$ to insert into Eq. (3.33).

The mass-shell constraint for a particle with mass $m$ is now given by

$$g_{\mu\nu} P^\mu P^\nu = -(1 + 2\Psi)(P^0)^2 + p^2 = -m^2, \tag{3.57}$$

where again

$$p^2 \equiv g_{ij} P^i P^j. \tag{3.58}$$

We will continue to define the energy as $E(p) \equiv \sqrt{p^2 + m^2}$. In the massless case, we obviously have $E = p$. We can now eliminate the time component of $P^\mu$ through

$$P^0 = \frac{E}{\sqrt{1 + 2\Psi}} = E(1 - \Psi). \tag{3.59}$$

This last equality holds since we are doing first-order perturbation theory in the small quantity $\Psi$. Similarly, we can use Eq. (3.58) to derive $P^i$. This yields the four-momentum of a massive particle in a perturbed FLRW spacetime (which includes the massless case):

$$P^\mu = \left[ E(1-\Psi),\, p^i \frac{1-\Phi}{a} \right]. \tag{3.60}$$

Here, we have defined $p^i$ through

$$p^i = p\,\hat{p}^i \quad \text{where} \quad \hat{p}^i = \hat{p}_i \tag{3.61}$$

is a unit vector satisfying $\delta_{ij}\hat{p}^i\hat{p}^j = 1$ as before. Eq. (3.60) allows us to eliminate $P^0$ and $P^i$ in favor of $E(p)$, $p$, the magnitude of the momentum, and $\hat{p}^i$ whenever they occur. Moreover, plugging these into Eq. (3.20) yields the expressions for the energy-momentum tensor in terms of the distribution function in the presence of metric perturbations (see Exercise 3.12) which we will need later.

Next, recall that (Eq. (3.26)) $P^i \equiv dx^i/d\lambda$ and $P^0 \equiv dt/d\lambda$. Therefore, at linear order in perturbations

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{dx^i}{d\lambda}\frac{d\lambda}{dt} \\ &= \frac{P^i}{P^0} = \frac{\hat{p}^i}{a}\frac{p}{E}(1-\Phi+\Psi). \end{aligned} \tag{3.62}$$

The remaining term to be calculated is $dp^i/dt$, from which we can deduce $dp/dt$ and $d\hat{p}^i/dt$. This proceeds via the geodesic equation just as in the homogeneous case in Sect. 3.2.2. The only difference is that we now have to deal with a larger number of Christoffel symbols due to the perturbations in the metric. While the calculation might appear cumbersome, it is straightforward, and the results will be useful in several different applications. In any case, we will only have to do it once! Throughout, we restrict to linear order in perturbations. To begin, let us evaluate the derivative of $p^i$ along the geodesic:

$$\begin{aligned} \frac{dp^i}{d\lambda} &= \frac{d}{d\lambda}\left[(1+\Phi)aP^i\right] \\ &= P^i\frac{d}{d\lambda}[(1+\Phi)a] + (1+\Phi)a\frac{dP^i}{d\lambda}. \end{aligned} \tag{3.63}$$

The first term can be computed using that $d/d\lambda = P^\mu\partial/\partial x^\mu$, which yields

$$\frac{d}{d\lambda}[(1+\Phi)a] = P^0 a[H + \dot{\Phi}] + aP^k\Phi_{,k}. \tag{3.64}$$

The second term can be evaluated using the geodesic equation:

$$\begin{aligned} \frac{dP^i}{d\lambda} &= -\Gamma^i{}_{\alpha\beta}P^\alpha P^\beta \\ &= -\left[\Gamma^i{}_{00}P^0P^0 + 2\Gamma^i{}_{0j}P^0P^j + \Gamma^i{}_{jk}P^jP^k\right]. \end{aligned} \tag{3.65}$$

Fortunately, the Christoffel symbols in the first and last terms here are already of first order, so that we can insert the zeroth-order expressions for $P^0$, $P^i$. The term $2\Gamma^i{}_{0j}P^0P^j$ contains the zeroth-order contribution. Using Eq. (3.56), we obtain

$$\frac{dP^i}{d\lambda} = -E\left\{\frac{E}{a^2}\Psi_{,i} + 2\left(H+\dot{\Phi}\right)\frac{p^i}{a}(1-\Psi-\Phi) + \frac{2}{a^2}\frac{p^i}{E}p^k\Phi_{,k} - \frac{p^2}{a^2E}\Phi_{,i}\right\}. \tag{3.66}$$

Eq. (3.63) thus becomes, via Eq. (3.60),

$$\begin{aligned}\frac{dp^i}{d\lambda} &= E(1-\Psi)\left\{[H+\dot{\Phi}]p^i + p^k\Phi_{,k}\frac{p^i}{aE}\right\} \\ &\quad - E\left\{\frac{E}{a}\Psi_{,i} + 2\left(H+\dot{\Phi}\right)p^i(1-\Psi) + \frac{2}{a}\frac{p^i}{E}p^k\Phi_{,k} - \frac{p^2}{aE}\Phi_{,i}\right\}.\end{aligned} \tag{3.67}$$

Finally, we can convert the total derivative with respect to $\lambda$ into a derivative with respect to $t$, using that $dp^i/dt = (P^0)^{-1}dp^i/d\lambda$:

$$\begin{aligned}\frac{dp^i}{dt} &= [H+\dot{\Phi}]p^i + p^k\Phi_{,k}\frac{p^i}{aE} \\ &\quad - \left\{\frac{E}{a}\Psi_{,i} + 2\left(H+\dot{\Phi}\right)p^i + \frac{2}{a}\frac{p^i}{E}p^k\Phi_{,k} - \frac{p^2}{aE}\Phi_{,i}\right\},\end{aligned} \tag{3.68}$$

which can be simplified to

$$\frac{dp^i}{dt} = -\left(H+\dot{\Phi}\right)p^i - \frac{E}{a}\Psi_{,i} - \frac{1}{a}\frac{p^i}{E}p^k\Phi_{,k} + \frac{p^2}{aE}\Phi_{,i}. \tag{3.69}$$

This is our desired result for the time evolution of the momentum $p^i$ along the geodesic. From now on, we will be able to always use the quantities $\{E, p, \hat{p}^i = \hat{p}_i\}$, i.e. the energy and physical momentum, rather than the comoving momentum $P^\mu$. Using that

$$\frac{dp}{dt} = \frac{d}{dt}\sqrt{\delta_{ij}p^ip^j} = \delta_{ij}\frac{p^i}{p}\frac{dp^j}{dt}, \tag{3.70}$$

for the magnitude of the momentum, we arrive at the corollary

$$\begin{aligned}\frac{dp}{dt} &= -\left[H+\dot{\Phi}\right]p - \frac{E}{a}\hat{p}^i\Psi_{,i} - \frac{1}{a}\frac{p^2}{E}\hat{p}^k\Phi_{,k} + \frac{p^2}{aE}\hat{p}^i\Phi_{,i} \\ &= -\left[H+\dot{\Phi}\right]p - \frac{E}{a}\hat{p}^i\Psi_{,i}.\end{aligned} \tag{3.71}$$

Eq. (3.71) describes the change in the magnitude of the momentum of a particle as it moves through a perturbed FLRW universe. While it took us quite some effort, we could have almost guessed its form without any calculation: the first term accounts for the loss of momentum due to the Hubble expansion (equivalent to the cosmological redshift and

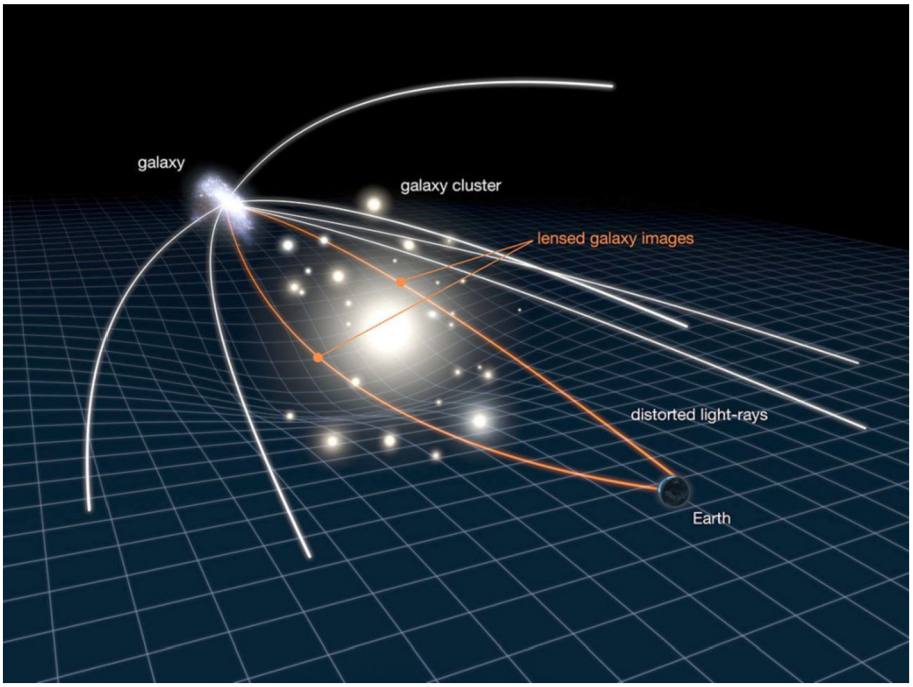



**FIGURE 3.4** Sketch illustrating how the curvature of spacetime induced by a massive cluster of galaxies deflects the trajectories of passing light rays. The same curvature also keeps the galaxies in orbit within the cluster. Both effects are described by Eq. (3.72). From www.cfhtlens.org.

decay of peculiar velocity). Recalling that we can interpret $\Phi$ as the fractional perturbation to the local scale factor, and that $H = \dot{a}/a$, we see that $H + \dot{\Phi}$ is the local expansion rate. Thus, the first two terms in Eq. (3.71) contain the cosmological redshift including its local perturbation. The last term describes the effect that a particle traveling into a well, such that $\hat{p}^i \partial\Psi/\partial x^i < 0$, gains energy; conversely, as it leaves the well, it loses energy. For nonrelativistic particles, this is familiar from Newtonian physics; it also holds for photons, however, in which case it corresponds to the gravitational redshift effect, which can be observed on Earth using precision atomic physics experiments.

Note that the terms involving $\Phi_{,i}$ and $p^k \Phi_{,k}$ in Eq. (3.69) have canceled in Eq. (3.71). They do not change the particle's momentum at linear order, but they do change the direction of the momentum. To see this, we can derive the time derivative of $\hat{p}^i$ from Eq. (3.69):

$$\begin{aligned} \frac{d\hat{p}^i}{dt} &= \frac{1}{p}\frac{dp^i}{dt} - \frac{p^i}{p^2}\frac{dp}{dt} \\ &= \frac{E}{ap}\left[\delta^{ik} - \hat{p}^i\hat{p}^k\right]\left(\frac{p^2}{E^2}\Phi - \Psi\right)_{,k} . \end{aligned} \tag{3.72}$$

Spatial gradients in the potentials change the trajectories of both massive and massless particles. Behind the mathematics, recall that the geodesic equation defines the notion of straight lines in curved space, where $\Phi$ and $\Psi$ capture the additional sources of curvature due to structure in the universe. This geometric content is illustrated in Fig. 3.4, which schematically shows a cluster of galaxies that distorts spacetime, leading to deflection of light rays passing by. It also deflects the motions of the galaxies within the cluster, which move in the same curved spacetime.

Consider first nonrelativistic particles with $p \ll E$. In this case, the term involving $\Phi$ is highly suppressed, so only $\Psi$, the perturbation to $g_{00}$, is relevant. Newtonian physics says that $d\boldsymbol{p}/dt = -m\nabla\Psi$, and you can check that Eq. (3.69) recovers this for nonrelativistic particles, up to an additional factor $1/a$. But this factor is also straightforward, since the spatial gradient we are taking is with respect to the comoving coordinate $x^k$, while Newton's force law refers to physical coordinates $a\boldsymbol{x}$. Next, consider photons. Then, $p/E = 1$ and both $\Phi$ and $\Psi$ contribute equally to the deflection. For $\Phi = -\Psi$, which we will see holds in many cases, this leads to the famous factor of 2 increase in the deflection of photons predicted by Einstein's theory over Newton and confirmed by measurements during the 1919 total solar eclipse. The effects described by Eq. (3.72) will become important when we study large-scale structure and gravitational lensing. Finally, notice that the deflection of nonrelativistic particles is much stronger than that of light. Indeed, photons are not bound to the cluster shown in Fig. 3.4 while matter and galaxies are. Mathematically, this is due to the factor $E/p$ in front which becomes large for nonrelativistic objects; physically, this happens because the curvature has more time to deflect the trajectories of slow-moving objects as compared to massless particles that travel at the speed of light.

With the geodesic equation, we have everything we need to write down the Boltzmann equation in the perturbed universe.

### 3.3.3 The collisionless Boltzmann equation for radiation

The Boltzmann equation for radiation, i.e. ultra-relativistic particles, in the perturbed universe is a straightforward generalization of the treatment in Sect. 3.2.2 which led us to Eq. (3.39). Moreover, we have done the hard part already by computing the expressions for $dx^i/dt$ [Eq. (3.62)] and $dp^i/dt$ [Eq. (3.69)]. We simply specialize them to the case $m = 0$, i.e. $E = p$. We can then write Eq. (3.33) as

$$\begin{aligned}\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x^i}\frac{\hat{p}^i}{a}(1 - \Phi + \Psi) - \frac{\partial f}{\partial p}\left\{\left[H + \dot{\Phi}\right]p + \frac{1}{a}p^i\Psi_{,i}\right\} \\ + \frac{\partial f}{\partial \hat{p}^i}\frac{1}{a}\left[(\Phi - \Psi)_{,i} - \hat{p}^i\hat{p}^k(\Phi - \Psi)_{,k}\right].\end{aligned} \tag{3.73}$$

This is the complete, linear-order left-hand side of the Boltzmann equation for radiation. However, we can simplify it further by making use of our knowledge of the zeroth-order distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$. In the homogeneous universe, this distribution is of the Bose–Einstein form Eq. (2.65). This equilibrium distribution obviously does not depend on

position $\boldsymbol{x}$, but it also does not depend on the direction of the momentum vector $\hat{\boldsymbol{p}}$ since it is isotropic. We now make the ansatz that the deviations from the equilibrium distribution of radiation in the inhomogeneous universe are of the same order as the spacetime perturbations $\Phi$, $\Psi$. We will see in subsequent chapters that this ansatz not only makes our life much easier, but is indeed valid.

With this working assumption, we can immediately drop the last term, $\propto \partial f/\partial \hat{p}^i$, in Eq. (3.73). Recall that $\partial f/\partial \hat{p}^i$ is nonzero only if we consider a perturbation to the zeroth order $f$; i.e., it is a first-order term. But so is the term which multiplies it. So we can neglect it.

Further, it is easy to see that the potentials in the second term $\propto \partial f/\partial x^i$ in Eq. (3.73) are higher order as well, because they multiply $\partial f/\partial x^i$ which is a first-order term (again, the zeroth-order distribution function does not depend on position). We finally obtain the Boltzmann equation for radiation consistently expanded to linear order:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - \left[H + \dot{\Phi} + \frac{1}{a}\hat{p}^i\frac{\partial \Psi}{\partial x^i}\right] p\frac{\partial f}{\partial p}. \tag{3.74}$$

Eq. (3.74) will lead us directly to the equations governing CMB anisotropies.

### 3.3.4 The collisionless Boltzmann equation for massive particles

We now perform the analogous derivation for particles that are not ultra-relativistic. This actually requires very little additional work. We start again from Eq. (3.33), and insert Eq. (3.62), Eq. (3.71), and Eq. (3.72):

$$\begin{aligned}\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x^i}\frac{\hat{p}^i}{a}\frac{p}{E}(1 - \Phi + \Psi) - p\frac{\partial f}{\partial p}\left[H + \dot{\Phi} + \frac{E}{ap}\hat{p}^i\Psi_{,i}\right] \\ + \frac{\partial f}{\partial \hat{p}^i}\frac{E}{ap}\left[\left(\frac{p^2}{E^2}\Phi - \Psi\right)_{,i} - \hat{p}^i\hat{p}^k\left(\frac{p^2}{E^2}\Phi - \Psi\right)_{,k}\right].\end{aligned} \tag{3.75}$$

We can now make the same assumptions about the zeroth-order distribution function of the massive particles as made for photons, namely that it is independent of position $\boldsymbol{x}$ and direction of the momentum vector $\hat{\boldsymbol{p}}$. This leads us to the linear-order Boltzmann equation for massive particles:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - \left[H + \dot{\Phi} + \frac{E}{ap}\hat{p}^i\Psi_{,i}\right] p\frac{\partial f}{\partial p}. \tag{3.76}$$

Eq. (3.76) reduces to Eq. (3.74) in the massless limit as it must. The main difference between the two is the presence of factors of $p/E$, or velocity, which by definition become unity for ultra-relativistic particles. For nonrelativistic matter on the other hand, the linear ansatz for the distribution function assumed here breaks down in the late universe. This will lead us to generalize the Boltzmann equation to nonlinear order in Ch. 12.

## 3.4 Summary

Almost all of cosmology consists of a series of applications of two fundamental equations of physics: the Einstein equations describing gravity; and the Boltzmann equation of statistical mechanics describing matter and radiation. In this chapter, we have provided a concise summary of these equations and applied them to the smooth and, in the case of the Boltzmann equation, perturbed universe.

The full ***Einstein equations*** are

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G\, T_{\mu\nu}, \tag{3.77}$$

where we have included the cosmological constant (or other form of dark energy) on the right-hand side. Applied to the FLRW metric and assuming a Euclidean universe, we derived the Friedmann equation for the scale factor $a(t)$:

$$\frac{H^2(t)}{H_0^2} = \frac{\rho(t)}{\rho_{\rm cr}} = \sum_{s={\rm r,m},\nu,{\rm DE}} \Omega_s \left[a(t)\right]^{-3(1+w_s)}. \tag{3.78}$$

Later chapters will be wholly devoted to studying perturbations around the homogeneous universe. Including these, we write the perturbed metric as

$$\begin{aligned} g_{00}(\boldsymbol{x}, t) &= -1 - 2\Psi(\boldsymbol{x}, t), \\ g_{0i}(\boldsymbol{x}, t) &= 0, \\ g_{ij}(\boldsymbol{x}, t) &= a^2(t)\delta_{ij}\left[1 + 2\Phi(\boldsymbol{x}, t)\right], \end{aligned} \tag{3.79}$$

and work to linear order in $\Psi$, $\Phi$ throughout. Deferring the derivation of the Einstein equations in the perturbed universe to Ch. 6, we solved the geodesic equation in the perturbed universe in this chapter. The comoving momentum becomes

$$P^\mu = \left[E(1-\Psi),\, p^i \frac{1-\Phi}{a}\right], \tag{3.80}$$

where $E = \sqrt{p^2 + m^2}$ is the proper energy and $\boldsymbol{p}$ is the physical momentum. The geodesic equation yields

$$\frac{dp^i}{dt} = -\left(H + \dot{\Phi}\right) p^i - \frac{E}{a}\Psi_{,i} - \frac{1}{a}\frac{p^i}{E} p^k \Phi_{,k} + \frac{p^2}{aE}\Phi_{,i}\,, \tag{3.81}$$

a compact relation which contains such diverse physics as Newtonian dynamics and gravitational lensing and which we will make use of many times throughout this book.

The ***Boltzmann equation*** involves two parts: the left-hand side is a total time derivative $df/dt$ expressing the conservation of the distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$ in the absence of collisions. This time derivative includes the effect of gravity as well. In the homogeneous

universe, the left-hand side becomes

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - Hp\frac{\partial f}{\partial p}. \tag{3.82}$$

Including structure in the universe via the perturbed metric Eq. (3.49), and assuming that all perturbations are small so that we can work to linear order in them, the Boltzmann equation becomes

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - \left[H + \dot{\Phi} + \frac{E}{ap}\hat{p}^i\Psi_{,i}\right]p\frac{\partial f}{\partial p}. \tag{3.83}$$

The second part of the Boltzmann equation is the collision term $C[f]$ on the right-hand side. It captures all microscopic scattering, pair production, annihilation, and decay processes. In particular, for a 2-particle scattering process

$$(1)_{\boldsymbol{p}} + (2)_{\boldsymbol{q}} \leftrightarrow (3)_{\boldsymbol{p}'} + (4)_{\boldsymbol{q}'}, \tag{3.84}$$

we derived the following collision term:

$$\begin{aligned} a\,C[f_1(\boldsymbol{p})] = {} & \frac{a}{2E_1(p)}\int\frac{d^3q}{(2\pi)^3 2E_2(q)}\int\frac{d^3p'}{(2\pi)^3 2E_3(p')}\int\frac{d^3q'}{(2\pi)^3 2E_4(q')}|\mathcal{M}|^2 \\ & \times (2\pi)^4\delta_{\rm D}^{(3)}[\boldsymbol{p}+\boldsymbol{q}-\boldsymbol{p}'-\boldsymbol{q}']\,\delta_{\rm D}^{(1)}[E_1(p)+E_2(q)-E_3(p')-E_4(q')] \\ & \times \Big\{f_3(\boldsymbol{p}')f_4(\boldsymbol{q}')\big[1\pm f_1(\boldsymbol{p})\big]\big[1\pm f_2(\boldsymbol{q})\big] \\ & - f_1(\boldsymbol{p})f_2(\boldsymbol{q})\big[1\pm f_3(\boldsymbol{p}')\big]\big[1\pm f_4(\boldsymbol{q}')\big]\Big\}. \end{aligned} \tag{3.85}$$

The distribution function in turn tells us what the ***energy-momentum tensor*** is that we have to put on the right-hand side of the Einstein equations. The general expression valid in the perturbed universe is Eq. (3.20). It looks more formidable than it is; as you can show in Exercise 3.12, Eq. (3.20) together with Eq. (3.80) yields the following energy-momentum tensor in the perturbed universe for a single species with degeneracy factor $g$:

$$\begin{aligned} T^0{}_0(\boldsymbol{x},t) &= -g\int\frac{d^3p}{(2\pi)^3}E(p)f(\boldsymbol{x},\boldsymbol{p},t), \\ T^0{}_i(\boldsymbol{x},t) &= g\,a(1+\Phi-\Psi)\int\frac{d^3p}{(2\pi)^3}p_i f(\boldsymbol{x},\boldsymbol{p},t), \\ T^i{}_j(\boldsymbol{x},t) &= g\int\frac{d^3p}{(2\pi)^3}\frac{p^i p_j}{E(p)}f(\boldsymbol{x},\boldsymbol{p},t). \end{aligned} \tag{3.86}$$

Notice that these simple expressions apply to the energy-momentum tensor with one raised index. In fact, the integral over $p_i f(\boldsymbol{x},\boldsymbol{p},t)$ in $T^0{}_i$ will turn out to be of first order in perturbations, so that we can drop the potentials $\Phi-\Psi$ in the prefactor as they lead to a second-order contribution.

## Exercises

**3.1** Calculate the curvature scalar, i.e. the 2D trace of the Ricci tensor, for the metric of 2D Euclidean space written in polar coordinates. Make use of the results of Exercise 2.2.

**3.2** Find the metric, Christoffel symbols, geodesic equation, and Ricci scalar for a $2+1$-dimensional spacetime given by the surface of a sphere with radius $r$.

**(a)** Using coordinates $t$, $\theta$, $\phi$, the metric is

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix}. \tag{3.87}$$

Show that the only nonvanishing Christoffel symbols are ${\Gamma^\theta}_{\phi\phi}$, ${\Gamma^\phi}_{\phi\theta}$, and ${\Gamma^\phi}_{\theta\phi}$. Express these in terms of $\theta$.

**(b)** Use these and the geodesic equation to find the equations of motion for a massive particle in this spacetime.

**(c)** Find the Ricci tensor. Show that contraction of this tensor leads to

$$R \equiv g^{\mu\nu} R_{\mu\nu} = \frac{2}{r^2}. \tag{3.88}$$

**3.3** Fill in some of the blanks left in our derivation of the Einstein equations.

**(a)** Compute the Christoffel symbol ${\Gamma^i}_{\alpha\beta}$ for a Euclidean FLRW metric.

**(b)** Compute the spatial components of the Ricci tensor in a Euclidean FLRW universe, $R_{ij}$. Show that the spacetime component, $R_{i0}$, vanishes.

**3.4** Show that the space-space component of the Einstein equations in a Euclidean universe is

$$\frac{\ddot{a}}{a} + \frac{1}{2}\left(\frac{\dot{a}}{a}\right)^2 = -4\pi G\mathcal{P} \tag{3.89}$$

where $\mathcal{P} = \delta^j_i T^i{}_j/3$ is the total pressure. Combine this with Eq. (3.12) to derive the *second Friedmann equation*:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left[\rho + 3\mathcal{P}\right]. \tag{3.90}$$

**3.5** Apply the Einstein equations to the case of an open universe. The spacetime interval in an open universe is[3]

$$ds^2 = -dt^2 + a^2(t)\left\{\frac{dr^2}{1+\Omega_K H_0^2 r^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right\} \tag{3.91}$$

where $r$, $\theta$, $\phi$ are 3D spherical coordinates, and $\Omega_K$ is the curvature parameter.

[3] This interval also describes a closed universe with $\Omega_K < 0$, but the coordinates chosen here then do not cover the entire universe, so we assume the universe to be open.

**(a)** First, compute the Christoffel symbols. Show that the only nonzero ones are equal to

$$\Gamma^{i}{}_{0j} = H\delta^{i}{}_{j}; \quad \Gamma^{0}{}_{ij} = g_{ij}H; \quad \Gamma^{i}{}_{jk} = \frac{g^{il}}{2}[g_{lj,k} + g_{lk,j} - g_{jk,l}]. \tag{3.92}$$

**(b)** Show that the components of the Ricci tensor are

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a}, \\ R_{ij} &= g_{ij}\left[\frac{\ddot{a}}{a} + 2H^2 - \frac{2\Omega_{\rm K}H_0^2}{a^2}\right]. \end{aligned} \tag{3.93}$$

**(c)** From these, compute the Ricci scalar, and then derive the time-time component of the Einstein equations. Compare with Eq. (3.14).

**3.6** By inserting Eq. (2.60) into Eq. (3.90), derive under what conditions the expansion of the universe is accelerating, assuming there is only a single constituent. What is the requirement for acceleration if there are multiple components $s$ with equations of state $w_s$?

**3.7** Eq. (3.20) gives the general-relativistic expression for the energy-momentum tensor in terms of the distribution function. Connect this to the expression Eq. (2.44) for the energy-momentum tensor in the homogeneous universe, with components given in Eq. (2.62) and Eq. (2.64).

**(a)** Using our results for $P^{\mu}$, derive $P_{\mu}$ and show that the spatial components $P_i$ are constant.

**(b)** Show that the time-time component of Eq. (3.20) agrees with the expression for the energy density given in Eq. (2.62).

**(c)** Show the same for the pressure given in Eq. (2.64).

**3.8** Derive the fluid equations for a collection of collisionless particles in a one-dimensional harmonic potential, by taking the *moments* of Eq. (3.23). That is, take the zeroth moment by integrating the equation over $dp/(2\pi)$ to obtain an equation that relates the number density $n$ to the fluid velocity $u$, where

$$n(x,t) \equiv \int_{-\infty}^{\infty} \frac{dp}{2\pi} f(x,p,t); \quad u(x,t) \equiv \frac{1}{n(x,t)} \int_{-\infty}^{\infty} \frac{dp}{2\pi} \frac{p}{m} f(x,p,t). \tag{3.94}$$

Then, take the first moment by multiplying the Boltzmann equation by $p$ and then integrating. Note that this equation governs the evolution of the fluid velocity $u$ and depends on the second moment, which contains the velocity dispersion. Neglect the velocity dispersion to close the set of equations, and make a mental note that this is quite general: when taking moments of the Boltzmann equation, the time evolution of each moment depends on a higher-order moment, so some approximation must be made in order to close the set of equations.

**3.9** Derive the time evolution of the zeroth-order distribution function for massive neutrinos, by using the zeroth-order Boltzmann equation (3.38) and assuming the Fermi–Dirac distribution as initial condition in the early universe (when neutrinos decoupled).

**(a)** Show that any distribution function of the form $f(p,t) = f(E_\nu[p_0 a(t)])$ solves the Boltzmann equation.

**(b)** By matching to the initial conditions at an early scale factor $a_{\rm dec}$, show that $p_0 = p/a_{\rm dec}$, and that the distribution function becomes

$$f_\nu^{(0)}(p,t) = f_{\rm FD}\left[E_\nu(a(t)\,p/a_{\rm dec})/T_{\rm dec}\right], \tag{3.95}$$

where $T_{\rm dec}$ is the temperature of the thermal neutrinos at $a_{\rm dec}$.

**(c)** Plot the neutrino distribution function at $z = 100,\ 10,\ 1$ and 0, for two cases: $m_\nu = 0.06$ eV and $m_\nu = 0$ (you can use the latter case to verify that you recover the correct result for massless neutrinos). Use $a_{\rm dec} = 10^{-9}$ and $T_{\rm dec} = T_{\nu,0}/a_{\rm dec}$, where $T_{\nu,0}$ is the extrapolated neutrino temperature today (see Eq. (2.81)).

**(d)** Show that, for this value of $a_{\rm dec}$, $T_{\rm dec} \gg m_\nu$ for realistic neutrino masses, and show that in this regime the distribution function matches Eq. (2.83). Here we have neglected any neutrino chemical potential, which is accurate as long as $\mu_\nu \ll T_{\rm dec}$.

**3.10** Derive Eq. (3.76), the Boltzmann equation for a massive particle at linear order, from Eq. (3.75).

**3.11** Show that the temperature of nonrelativistic matter scales as $a^{-2}$ in the absence of interactions. Start from Eq. (3.38) and assume a form $f_{\rm c} \propto e^{-(E-\mu)/T} \propto e^{-p^2/2mT}$. Justify this ansatz. Note that this argument does *not* apply in the presence of interactions: for example, the temperature of electrons and protons scales as $a^{-1}$ as long as they are tightly coupled to the photons.

**3.12** Derive Eq. (3.86), by using the perturbed metric Eq. (3.49), and inserting Eq. (3.60) into Eq. (3.20).

**3.13** Show that the spatial curvature in conformal Newtonian gauge is equal to $4k^2\Phi/a^2$. To do this, compute the three-dimensional Ricci scalar arising from the spatial part of the metric $g_{ij}$ in Eq. (3.49).

# 4

# The origin of species

The very early universe was hot and dense. As a result, interactions among particles occurred much more frequently than they do today. As an example, a photon in the visible band today can typically travel across much of the observable universe without deflection or capture, so it has a mean free path greater than $10^{28}$ cm. When the age of the universe was equal to 1 sec, though, the mean free path of a photon was about the size of an atom. Thus, in the time it took the universe to expand by a factor of 2, a given photon interacted many, many times. These multiple interactions kept many of the constituents in the universe in equilibrium. Nonetheless, there were times when reactions could not proceed rapidly enough to maintain equilibrium conditions. Not coincidentally, these times are of the utmost interest to cosmologists.

Indeed, we will see in this chapter that out-of-equilibrium phenomena played a role in $(i)$ the formation of the light elements during Big Bang Nucleosynthesis; $(ii)$ recombination of electrons and protons into neutral hydrogen; and possibly in $(iii)$ the production of dark matter in the early universe. It is important to understand that all three phenomena are the result of nonequilibrium physics and that all three can be studied with the same formalism: the Boltzmann equation in the homogeneous universe, as introduced in Sect. 3.2. Sects. 4.2–4.4 of this chapter are simply applications of this general formula.

To summarize, in this chapter we will go beyond our treatment in Ch. 2 by considering out-of-equilibrium processes in the universe, but we still work within the framework of a homogeneous universe. In succeeding chapters, we will then move beyond uniformity and explore distribution functions for matter and radiation that depend on both position and direction of propagation.

## 4.1 The homogeneous Boltzmann equation revisited

Suppose that we are interested in the number density $n_1$ of species 1. For simplicity, we will assume that the only process affecting the abundance of this species is a reaction with species 2 producing two particles, imaginatively called 3 and 4. Schematically, $1 + 2 \leftrightarrow 3 + 4$; i.e., particle 1 and particle 2 can annihilate producing particles 3 and 4, or the inverse process can produce 1 and 2. The Boltzmann equation for this system in an expanding universe was derived in Sect. 3.2.2, and the corresponding collision term in Sect. 3.2.3. Combining the general results Eq. (3.43) and Eq. (3.48), we obtain the following evolution equation for $n_1$:

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00010-3
Copyright © 2021 Elsevier Inc. All rights reserved.

$$a^{-3}\frac{d(n_1a^3)}{dt} = \int \frac{d^3p_1}{(2\pi)^3 2E_1}\int \frac{d^3p_2}{(2\pi)^3 2E_2}\int \frac{d^3p_3}{(2\pi)^3 2E_3}\int \frac{d^3p_4}{(2\pi)^3 2E_4}$$
$$\times (2\pi)^4\delta_{\rm D}^{(3)}(\boldsymbol{p}_1+\boldsymbol{p}_2-\boldsymbol{p}_3-\boldsymbol{p}_4)\delta_{\rm D}^{(1)}(E_1+E_2-E_3-E_4)|\mathcal{M}|^2$$
$$\times \{f_3f_4[1\pm f_1][1\pm f_2]-f_1f_2[1\pm f_3][1\pm f_4]\}. \tag{4.1}$$

Here, $E_i$ stands for $E_i(p_i)$ and $f_i$ for $f_i(p_i,t)$. We have thus obtained an integrodifferential equation for the phase-space distributions. Further, in principle at least, it must be supplemented with similar equations for the other species. In practice, these formidable obstacles can be overcome for many practical cosmological applications. The first, most important, realization is that scattering processes typically enforce *kinetic equilibrium*. That is, scattering takes place so rapidly that the distributions of the various species take on the generic Bose–Einstein/Fermi–Dirac forms (Eq. (2.65) and Eq. (2.66)) with equal temperature $T$ for each species. This form condenses all of the freedom in the distribution into the functions of time $T$ and $\mu$. If annihilations were also in equilibrium, the sum of the chemical potentials $\mu_i$ in any reaction would have to balance. For example, the reaction $e^+ + e^- \leftrightarrow \gamma+\gamma$ would lead to $\mu_{e^+}+\mu_{e^-}=2\mu_\gamma$. In the out-of-equilibrium cases we will study, the system will not be in *chemical equilibrium* and we will have to solve a differential equation for $\mu$. The great simplifying feature of kinetic equilibrium, though, is that this differential equation will be a single ordinary differential equation, as opposed to the very complicated form of Eq. (4.1).

We will typically be interested in systems at temperatures smaller than $E-\mu$. In this limit, the exponential in the Bose–Einstein and Fermi–Dirac distributions is large and dwarfs the $\pm1$ in the denominator. Thus, another simplification emerges: we can ignore the complications of quantum statistics, and the distributions follow the *Boltzmann distribution* of a classical dilute gas:

$$f(E)\to e^{\mu/T}e^{-E/T}. \tag{4.2}$$

In addition, the Pauli blocking/Bose enhancement factors in the Boltzmann equation can be neglected. This is because $f(E)\ll 1$ in the limit of Eq. (4.2).

Under these approximations, the last line of Eq. (4.1) becomes

$$f_3f_4[1\pm f_1][1\pm f_2]-f_1f_2[1\pm f_3][1\pm f_4]$$
$$\to\quad e^{-(E_1+E_2)/T}\left\{e^{(\mu_3+\mu_4)/T}-e^{(\mu_1+\mu_2)/T}\right\}. \tag{4.3}$$

Here we have used energy conservation, $E_1+E_2=E_3+E_4$. We will use the number densities themselves as the time-dependent functions to be solved for, instead of $\mu$. The number density of species $s$ is related to $\mu_s$ via

$$n_s = g_s e^{\mu_s/T}\int \frac{d^3p}{(2\pi)^3}e^{-E_s(p)/T}, \tag{4.4}$$

where $g_s$ is the now familiar degeneracy factor of the species. It will also be useful to define the species-dependent number density for $\mu_s = 0$ as

$$n_s^{(0)} \equiv g_s \int \frac{d^3p}{(2\pi)^3} e^{-E_s(p)/T} = \begin{cases} g_s \left(\frac{m_s T}{2\pi}\right)^{3/2} e^{-m_s/T} & m_s \gg T, \\ g_s \frac{T^3}{\pi^2} & m_s \ll T. \end{cases} \tag{4.5}$$

In particular, we have the very useful relation $n_\gamma^{(0)} = 2T^3/\pi^2$. With this definition, $e^{\mu_i/T}$ can be rewritten as $n_i/n_i^{(0)}$, so via Eq. (4.4) we have

$$e^{(\mu_3+\mu_4)/T} - e^{(\mu_1+\mu_2)/T} = \frac{n_3 n_4}{n_3^{(0)} n_4^{(0)}} - \frac{n_1 n_2}{n_1^{(0)} n_2^{(0)}}. \tag{4.6}$$

With these approximations, the Boltzmann equation now simplifies enormously. Define the thermally averaged cross section as

$$\begin{aligned} \langle \sigma v \rangle \equiv \frac{1}{n_1^{(0)} n_2^{(0)}} \int \frac{d^3p_1}{(2\pi)^3 2E_1} \int \frac{d^3p_2}{(2\pi)^3 2E_2} \int \frac{d^3p_3}{(2\pi)^3 2E_3} \int \frac{d^3p_4}{(2\pi)^3 2E_4} e^{-(E_1+E_2)/T} \\ \times (2\pi)^4 \delta_{\rm D}^{(3)}(\boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_3 - \boldsymbol{p}_4) \delta_{\rm D}^{(1)}(E_1 + E_2 - E_3 - E_4) |\mathcal{M}|^2, \end{aligned} \tag{4.7}$$

which in general depends on the temperature $T$. Then, the Boltzmann equation becomes

$$a^{-3} \frac{d(n_1 a^3)}{dt} = n_1^{(0)} n_2^{(0)} \langle \sigma v \rangle \left\{ \frac{n_3 n_4}{n_3^{(0)} n_4^{(0)}} - \frac{n_1 n_2}{n_1^{(0)} n_2^{(0)}} \right\}. \tag{4.8}$$

We thus have a simple ordinary differential equation for the number density of each species. Although the details will vary from application to application (see Table 4.1), we will always start from this equation when tracking abundances.

One qualitative note about Eq. (4.8). The left-hand side is of order $n_1/t$, or, since the typical cosmological time is $H^{-1}$, $n_1 H$. The right-hand side is of order $n_1 n_2 \langle \sigma v \rangle$. Therefore, if the reaction rate for a single particle of type 1, $n_2 \langle \sigma v \rangle$, is much larger than the expansion rate, then the terms on the right side will be much larger than the one on the left. The only way to maintain equality then is for the individual terms on the right to cancel. Thus, when reaction rates are large, Eq. (4.8) approaches

$$\frac{n_3 n_4}{n_3^{(0)} n_4^{(0)}} = \frac{n_1 n_2}{n_1^{(0)} n_2^{(0)}}. \tag{4.9}$$

This equation is equivalent to the condition $\mu_1 + \mu_2 = \mu_3 + \mu_4$, the relation we have referred to as *chemical equilibrium* above. The same relation is also known under the names of *nuclear statistical equilibrium* and *Saha equation*.

**Table 4.1** The most important reactions discussed in this Chapter. In the last row, $X$ denotes a dark matter particle, while $\psi$ denotes lighter particles produced upon annihilation.

| | 1 | 2 | $\leftrightarrow$ | 3 | 4 |
|---|---|---|---|---|---|
| Neutron–proton ratio | $n$ | $\nu_e$ or $e^+$ | $\leftrightarrow$ | $p$ | $e^-$ or $\overline{\nu}_e$ |
| Recombination | $e$ | $p$ | $\leftrightarrow$ | H | $\gamma$ |
| Dark matter production | $X$ | $X$ | $\leftrightarrow$ | $\psi$ | $\psi$ |

## 4.2 Big Bang nucleosynthesis

Of the various epochs in the early universe, we have seen in Ch. 1 that Big Bang Nucleosynthesis (BBN) is of particular importance, as it produced the light elements we see in the universe and can be used to constrain cosmology. BBN happened when the temperature of the universe cooled to 1 MeV. At that point in time, the cosmic plasma consisted of:

- **Relativistic particles in equilibrium: photons, electrons and positrons.** These were kept in close contact with each other by electromagnetic interactions such as $e^+e^- \leftrightarrow \gamma\gamma$. Besides a small difference due to fermion/boson statistics, these all had the same abundances.
- **Decoupled relativistic particles: neutrinos.** At temperatures a little above 1 MeV, the rate for processes such as $\nu e \leftrightarrow \nu e$ that keep neutrinos coupled to the rest of the plasma dropped beneath the expansion rate. Neutrinos therefore share the same temperature as the other relativistic particles (but see Sect. 2.4.4), and hence are roughly as abundant, but they do not couple to them.
- **Nonrelativistic particles: baryons.** If there had been no asymmetry in the initial number of baryons and anti-baryons, then both would be completely depleted by 1 MeV. However, such an asymmetry has to exist, since otherwise we would observe a universe almost completely devoid of baryons. Comparing the abundance of baryons to photons, we find $n_\mathrm{b}/s \sim 10^{-10}$ today.[1] Since this ratio remains constant throughout the expansion (as long as the baryon number density is conserved), this also quantifies the baryon–antibaryon asymmetry in the early universe. As you can compute in Exercise 4.6,

$$\eta_\mathrm{b} \equiv \frac{n_\mathrm{b}}{n_\gamma} = 6.0 \times 10^{-10} \left(\frac{\Omega_\mathrm{b} h^2}{0.022}\right). \tag{4.10}$$

  There are thus many fewer baryons than relativistic particles in the universe.

Our task in this section will be to determine what nuclei the protons and neutrons end up in, and in which amounts. Were the system to remain in equilibrium throughout, the final state would be dictated solely by energetics, and all baryons would relax to the nuclear

[1] $s$ is the entropy density, which scales as $a^{-3}$, as we saw in Ch. 2.

state with the lowest energy per baryon, iron (Fig. 4.1). However, nuclear reactions, whose rates scale as the second—or higher—power of the density, are too slow to keep the system in equilibrium as the universe expands. In principle then, we need to solve the equivalent of Eq. (4.8) for all the nuclei, i.e., a set of coupled differential equations. In practice, at least for a qualitative understanding of the result, we can make use of two simplifications that obviate the need to solve the full set of differential equations.

The first simplification is that essentially no elements heavier than helium are produced at appreciable levels. So the only nuclei that we need to trace are hydrogen and helium, and their isotopes: deuterium, tritium, and $^3$He. The second simplification is that, even in the context of this reduced set of elements, the physics splits up neatly into two parts since above $T \simeq 0.1$ MeV, no light nuclei form: only free protons and neutrons exist. Therefore, we first solve for the neutron/proton ratio and then use this abundance as input for the synthesis of helium and isotopes such as deuterium (see Box 4.1).

### 4.1 Lightning Introduction to Nuclear Physics

Atomic nuclei are characterized by two numbers: the *atomic number* $Z$ gives the number of protons in the nucleus (and hence its charge), and is unique to each element; the *mass number* $A$ is the total number of neutrons and protons in the nucleus. Nuclei with different $A$ but the same $Z$ are referred to as *isotopes*. The mass number is denoted by a superscript before the name of the element. So, a single proton $p$ can be equivalently written as the hydrogen nucleus $^1$H. A deuterium nucleus consists of a proton and a neutron and is written as $^2$H or D; one proton and two neutrons make tritium, $^3$H. Nuclei with $Z = 2$ are helium; these can have one neutron ($^3$He) or two ($^4$He).

The total mass of a nucleus with $Z$ protons and $A - Z$ neutrons differs slightly from the sum of masses of the individual protons and neutrons. This difference is called the binding energy, which is defined as

$$B_{\rm N} \equiv Z m_p + (A - Z) m_n - m_{\rm N} \tag{4.11}$$

where $m_{\rm N}$ is the mass of the nucleus. For example, the mass of deuterium is 1875.62 MeV while the sum of the neutron and proton masses is 1877.84 MeV, so the binding energy of deuterium is $B_{\rm D} = 2.22$ MeV. Nuclear binding energies are typically in the MeV range, which explains why Big Bang Nucleosynthesis occurs at temperatures a bit lower than 1 MeV even though nuclear masses are in the GeV range.

Neutrons and protons can interconvert via weak interactions:

$$p + \bar{\nu} \leftrightarrow n + e^+;\ p + e^- \leftrightarrow n + \nu;\ n \leftrightarrow p + e^- + \bar{\nu} \tag{4.12}$$

where all the reactions can proceed in either direction. The light elements are built up via nuclear interactions. For example, deuterium forms from $p + n \rightarrow$ D. Then, D + D $\rightarrow n + {}^3$He, after which $^3$He + D $\rightarrow p + {}^4$He produces $^4$He. Here the final-state nuclei are usually in an excited state, and then relax to the ground state by emitting one or more photons.

Both of these simplifications—no heavy elements at all and only $n/p$ above 0.1 MeV—rely on the physical fact that, at high temperatures, comparable to nuclear binding energies, any time a nucleus is produced in a reaction, it is destroyed by a high-energy photon.

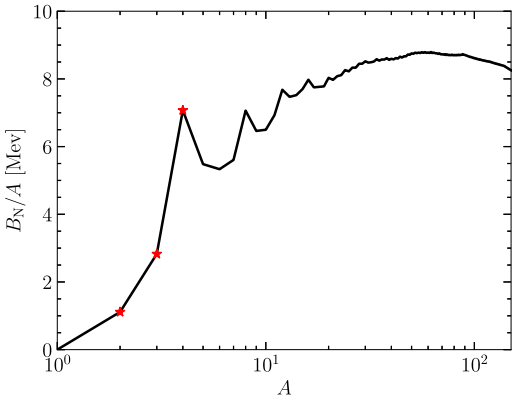



**FIGURE 4.1** Binding energy per nucleon as a function of mass number $A$ of nuclei (data are from Audi et al., 2003). The stars indicate the three isotopes most relevant for BBN: D ($A = 2$), $^3$He ($A = 3$), and $^4$He ($A = 4$). Among the light elements, $^4$He is a crucial local maximum. Nucleosynthesis in the early universe essentially stops at $^4$He because of the lack of tightly bound isotopes at $A = 5 - 7$. In the high-density environment of stars, three $^4$He nuclei fuse to form $^{12}$C, but the low baryon number precludes this process in the early universe.

This fact is reflected in the fundamental equilibrium equation (4.9). To see how, let us consider this equation applied to deuterium production, $n + p \leftrightarrow \mathrm{D} + \gamma$. Since photons have $n_\gamma = n_\gamma^{(0)}$ (the chemical potential $\mu_\gamma$ is extremely small), the equilibrium condition becomes

$$\frac{n_\mathrm{D}}{n_n n_p} = \frac{n_\mathrm{D}^{(0)}}{n_n^{(0)} n_p^{(0)}}. \tag{4.13}$$

Using Eq. (4.5) for the quantities on the right-hand side leads to

$$\frac{n_\mathrm{D}}{n_n n_p} = \frac{3}{4}\left(\frac{2\pi m_\mathrm{D}}{m_n m_p T}\right)^{3/2} e^{[m_n + m_p - m_\mathrm{D}]/T}, \tag{4.14}$$

the factor of 3/4 being due to the number of spin states (3 for D and 2 each for $p$ and $n$). In the prefactor, $m_\mathrm{D}$ can be set to $2m_n = 2m_p$, but in the exponential the small difference between $m_n + m_p$ and $m_\mathrm{D}$ is important: indeed the argument of the exponential is equal to the binding energy of deuterium, $B_\mathrm{D} = 2.22$ MeV (see Box 4.1), divided by the temperature. Therefore, as long as equilibrium holds,

$$\frac{n_\mathrm{D}}{n_n n_p} = \frac{3}{4}\left(\frac{4\pi}{m_p T}\right)^{3/2} e^{B_\mathrm{D}/T}. \tag{4.15}$$

Both the neutron and proton density are proportional to the baryon density, so using $n_n \simeq n_p \simeq n_\mathrm{b} = \eta_\mathrm{b} n_\gamma^{(0)}$ we roughly have

$$\frac{n_\mathrm{D}}{n_\mathrm{b}} \sim \eta_\mathrm{b}\left(\frac{T}{m_p}\right)^{3/2} e^{B_\mathrm{D}/T}. \tag{4.16}$$

As long as $B_D/T$ is not too large, the small prefactor dominates this expression—recall the smallness of the baryon-to-photon ratio $\eta_b$, Eq. (4.10).

The small baryon-to-photon ratio thus inhibits nuclei production until the temperature drops well beneath the nuclear binding energy. Physically, this happens because there are so many photons around that, even though photons with energies of order the nuclear binding energy are in the exponentially suppressed tail of the Bose–Einstein distribution (since $T \ll B_D$), any given nucleus will still encounter at least one such photon within a Hubble time until the temperature drops even further. So, at temperatures above 0.1 MeV, virtually all baryons are in the form of neutrons and protons. When the temperature falls to 0.1 MeV, roughly, deuterium and helium are produced, but the reaction rates are by now too low to produce any heavier elements. Fig. 4.1 explains this. The lack of a stable isotope with mass number 5 implies that heavier elements cannot be produced via $^4\mathrm{He} + p \to X$. In stars, the triple-alpha process $^4\mathrm{He} + {}^4\mathrm{He} + {}^4\mathrm{He} \to {}^{12}\mathrm{C}$ produces heavier elements, but in the early universe, by the time that $^4$He can form the densities are far too low to allow three nuclei to find one another within a Hubble time.

### 4.2.1 Neutron abundance

We begin by solving for the neutron–proton ratio. Protons can be converted into neutrons via weak interactions, $p + e^- \to n + \nu_e$ for example. As we will see, reactions of this sort keep neutrons and protons in equilibrium until $T \sim$ MeV. Thereafter, one must solve Eq. (4.8) to track the neutron abundance.

From Eq. (4.5), the proton/neutron equilibrium ratio in the nonrelativistic limit (so that $E_i(p) = m_i + p^2/2m_i$) is

$$\frac{n_p^{(0)}}{n_n^{(0)}} = \frac{e^{-m_p/T}\int dp\, p^2 e^{-p^2/2m_pT}}{e^{-m_n/T}\int dp\, p^2 e^{-p^2/2m_nT}}. \tag{4.17}$$

The integrals here are proportional to $m^{3/2}$. The resulting ratio $(m_p/m_n)^{3/2}$ is sufficiently close to unity that we can neglect the mass difference. However, in the exponential the mass difference is very important, and we are left with

$$\frac{n_p^{(0)}}{n_n^{(0)}} = e^{\mathcal{Q}/T} \tag{4.18}$$

with $\mathcal{Q} \equiv m_n - m_p = 1.293$ MeV. Therefore, at high temperatures, there are as many neutrons as protons. As the temperature drops beneath 1 MeV, the neutron fraction goes down. If weak interactions operated efficiently enough to maintain equilibrium indefinitely, then it would drop to zero (even if free neutrons were stable). The main task of this section is to find what happens in the real world where weak interactions are not so efficient. It is convenient to define

$$X_n \equiv \frac{n_n}{n_n + n_p}, \tag{4.19}$$

that is, $X_n$ is the ratio of neutrons to total nuclei. In equilibrium,

$$X_n \to X_{n,\mathrm{EQ}} \equiv \frac{1}{1 + n_p^{(0)}/n_n^{(0)}}. \tag{4.20}$$

To track the evolution of $X_n$, let us start from Eq. (4.8), with 1 = neutron, 3 = proton, and 2, 4 = leptons in complete equilibrium ($n_l = n_l^{(0)}$). Then,

$$a^{-3}\frac{d(n_n a^3)}{dt} = n_l^{(0)}\langle\sigma v\rangle\left\{\frac{n_p n_n^{(0)}}{n_p^{(0)}} - n_n\right\}. \tag{4.21}$$

We have already determined the ratio $n_n^{(0)}/n_p^{(0)} = e^{-\mathcal{Q}/T}$ and we can identify $n_l^{(0)}\langle\sigma v\rangle$ as $\lambda_{np}$, the rate for neutron → proton conversion since it multiplies $n_n$ in the loss term. Also, if we rewrite $n_n$ on the left as $(n_n + n_p)X_n$, then the total density times $a^3$ can be taken outside the derivative, leaving

$$\frac{dX_n}{dt} = \lambda_{np}\left\{(1 - X_n)e^{-\mathcal{Q}/T} - X_n\right\}. \tag{4.22}$$

Eq. (4.22) is a differential equation for $X_n$ as a function of time. It turns out that it is simpler to solve once we recast the equation using a new evolution variable $x$,

$$x \equiv \frac{\mathcal{Q}}{T}. \tag{4.23}$$

The left-hand side of Eq. (4.22) then becomes $\dot{x}\, dX_n/dx$, so we need an expression for $\dot{x} = -x\dot{T}/T$. Since $T \propto a^{-1}$,

$$\frac{1}{T}\frac{dT}{dt} = -H = -\sqrt{\frac{8\pi G\rho}{3}}, \tag{4.24}$$

with the second equality following from Eq. (3.12). Nucleosynthesis occurs in the radiation-dominated era, so the main contribution to the energy density $\rho$ comes from relativistic particles. Recall from Ch. 2 that the contribution to the energy density from relativistic particles is

$$\begin{aligned}\rho &= \frac{\pi^2}{30}T^4\left[\sum_{s=\mathrm{bosons}} g_s + \frac{7}{8}\sum_{s=\mathrm{fermions}} g_s\right] \quad (s \text{ relativistic}) \\ &\equiv g_*\frac{\pi^2}{30}T^4.\end{aligned} \tag{4.25}$$

The effective numbers of relativistic degrees of freedom, $g_*$, is a function of the temperature. At temperatures of order 1 MeV, the contributing species are: photons ($g_\gamma = 2$), neutrinos ($g_\nu = 6$), and electrons and positrons ($g_{e^+} = g_{e^-} = 2$). Adding up leads to $g_* \simeq 10.75$, roughly constant throughout the regime of interest. Then, Eq. (4.22) becomes

$$\frac{dX_n}{dx} = \frac{x\lambda_{np}}{H(x=1)}\left\{e^{-x} - X_n(1 + e^{-x})\right\}. \tag{4.26}$$

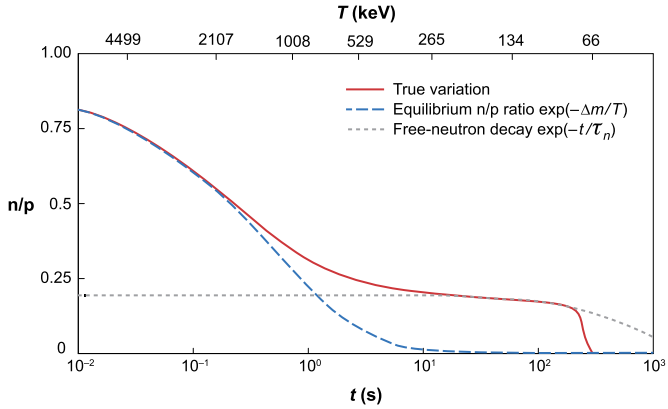



**FIGURE 4.2** Evolution of the neutron-to-proton ratio $n_n/n_p = X_n/(1-X_n)$ in the early universe. The solid curve shows the full numerical result. The long-dashed curve shows the equilibrium prediction (with $\Delta m \equiv \mathcal{Q}$), while the short-dashed curve shows the decay factor $\exp(-t/\tau_n)$. The neutron abundance falls out of equilibrium at $T \sim 1$ MeV, and BBN sets in at $T \sim 0.1$ MeV, leading to the sharp drop in the neutron abundance. From Steigman (2007).

In Exercise 4.5, you will integrate this equation numerically to track the neutron abundance. It turns out that when $T = \mathcal{Q}$ (i.e., when $x = 1$), the conversion rate is 5.5 $s^{-1}$, somewhat larger than the expansion rate. As the temperature drops beneath 1 MeV, though, the reaction rate falls as $T^3$, while the expansion rate falls as $T^2$, so conversions become inefficient, and we expect a departure from the equilibrium result.

Fig. 4.2 shows the evolution of $X_n$ (in fact, a more precise calculation which includes proper statistics, nonzero electron mass, and changing $g_*$). The neutron fraction $X_n$ does indeed fall out of equilibrium once the temperature drops below 1 MeV: it freezes out at roughly 0.15 (corresponding to the $y$-axis ratio equal to 0.18) once the temperature drops below 0.5 MeV. Here, the term "freezing out" or "freeze-out" means the inability of annihilations to keep the particle's abundance at its equilibrium value. At temperatures below 0.1 MeV, two reactions we have not included yet become important: neutron decay ($n \to p + e^- + \overline{\nu}$) and deuterium production ($n + p \to \mathrm{D} + \gamma$), i.e. the beginning of BBN. Deuterium production, which we will study in the next section, leads to the sharp drop at $T \simeq 0.1$ MeV.

First, decays can be added trivially by multiplying the neutron abundance with a factor of $e^{-t/\tau_n}$, where the neutron lifetime is $\tau_n = (885.7 \pm 0.8)$ s. By the time decays become important, electrons and positrons have annihilated, so $g_*$ in Eq. (4.25) is 3.36 and the time-temperature relation is (Exercise 2.5):

$$t = 132\,\mathrm{s}\left(\frac{0.1\ \mathrm{MeV}}{T}\right)^2. \tag{4.27}$$

We will see shortly that production of deuterium, and other light elements, begins in earnest at $T_{\mathrm{nuc}} \sim 0.07$ MeV. By then, decays have depleted the neutron fraction by a factor

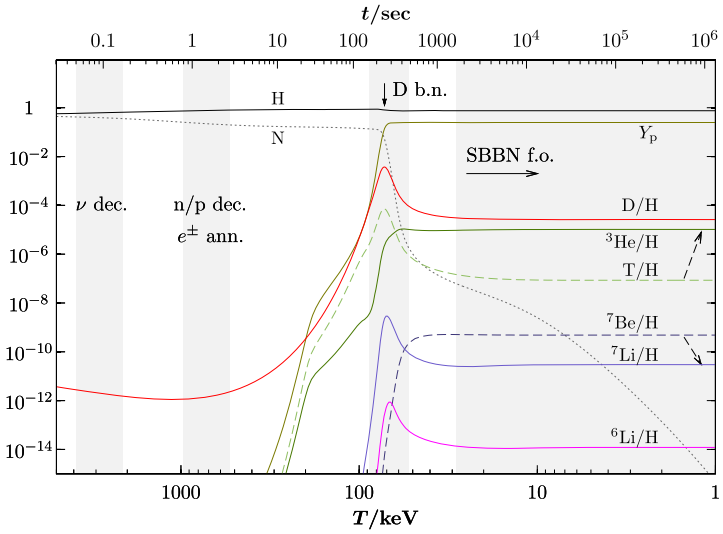



**FIGURE 4.3** Evolution of the mass fraction in light elements during BBN (SBBN stands for "standard BBN"). The lower $x$ axis shows temperature, while the upper $x$ axis shows time. The abundance of deuterium peaks during BBN and then decays as deuterium gets processed to helium as well as trace amounts of other elements. From Pospelov and Pradler (2010).

of $\exp[-(132/886)(0.1/0.07)^2] = 0.74$. So the neutron abundance at the onset of nucleosynthesis is $0.15 \times 0.74$, or

$$X_n(T_{\rm nuc}) = 0.11. \tag{4.28}$$

We now turn to light element formation to understand the ramifications of this number.

## 4.2.2 Light element abundances

A useful way to approximate light element production is that it occurs instantaneously at a temperature $T_{\rm nuc}$ when the energetics compensates for the small baryon-to-photon ratio. Let us consider deuterium production as an example, with Eq. (4.16) as our guide. The equilibrium deuterium abundance is of order the baryon abundance (i.e. if the universe stayed in equilibrium, all neutrons and protons would form deuterium) when the right-hand side of Eq. (4.16) is of order unity, or

$$\ln(\eta_{\rm b}) + \frac{3}{2}\ln(T_{\rm nuc}/m_p) \sim -\frac{B_{\rm D}}{T_{\rm nuc}}. \tag{4.29}$$

Eq. (4.29) suggests that deuterium production takes place at $T_{\rm nuc} \sim 0.07$ MeV.

Since the binding energy of helium is larger than that of deuterium, the exponential factor $e^{B/T}$ favors helium over deuterium. Indeed, Fig. 4.3 illustrates that helium is produced almost immediately after deuterium. Virtually all remaining neutrons at $T \sim T_{\rm nuc}$ then are processed into $^4$He. Since two neutrons go into $^4$He, the final $^4$He abundance is equal to

half the neutron abundance at $T_{\rm nuc}$. Often, results are quoted in terms of mass fraction; then,

$$Y_P \equiv \frac{4n(^4{\rm He})}{n_{\rm b}} = 2X_n(T_{\rm nuc}), \tag{4.30}$$

which yields a final helium mass fraction of 0.22. This rough estimate, obtained by solving a single differential equation, is in remarkable agreement with a full numerical calculation, which can be fit via (Olive, 2000)

$$Y_P = 0.2262 + 0.0135\ln(\eta_{\rm b}/10^{-10}). \tag{4.31}$$

One important feature of this result is that it depends only logarithmically on the baryon fraction via $\eta_{\rm b}$, which is inherited from Eq. (4.29). You might think that the exponential sensitivity to $T_{\rm nuc}$ in the decay fraction would turn this into a linear dependence. However, $T_{\rm nuc}$ is sufficiently early that only a small fraction of neutrons have decayed: the exponential in this regime is linear in the time. Therefore, the final helium abundance maintains only a logarithmic dependence on the baryon density. This weak dependence on $\Omega_{\rm b}h^2$ is clearly visible in Fig. 1.6, which also shows that the prediction agrees well with the observations (horizontal shaded band). The best indication of the primordial helium abundance comes from the most unprocessed gas, indicated by a elemental composition that is close to the primordial one, i.e. elements heavier than Helium are virtually absent.

Fig. 4.3 shows that not all of the deuterium gets processed into helium. A trace amount remains unburned, simply because the reaction that eliminates it, ${\rm D} + p \to {}^3{\rm He} + \gamma$, is not completely efficient. While deuterium is depleted via these reactions after $T_{\rm nuc}$, it eventually freezes out at a mass fraction of order $3 \times 10^{-5}$. If the baryon density is low, then the reactions proceed more slowly, and the depletion is not as effective. Therefore, low baryon density inevitably results in more deuterium; the sensitivity is quite stark, as illustrated in Fig. 1.6. This fact, combined with the possibility of measuring deuterium in high-redshift gas clouds by looking for absorption in the spectra of distant QSOs (see Sect. 1.3), turns the deuterium abundance into an important probe of the baryon density.

## 4.3 Recombination

After BBN is complete, the ordinary matter in the universe consists of protons, electrons, photons, helium nuclei and trace amounts of heavier nuclei (the neutrinos have decoupled by now, and no longer play a role). The next important epoch is when the Compton scattering between photons and electrons is no longer efficient enough to keep the photons tightly coupled to the baryons (electrons and baryons remain tightly coupled by Coulomb scattering throughout). This *decoupling* epoch happens as the temperature drops below $\sim 1$ eV when the number of free electrons drops dramatically. When $T \geq 1$ eV, there is still very little neutral hydrogen. Energetics of course favor the production of neutral hydrogen with a binding energy of $\epsilon_0 = 13.6$ eV, but the high photon/baryon ratio, with the

correspondingly large number of photons in the high-energy tail of the Bose–Einstein distribution, ensures that any hydrogen atom produced will be immediately ionized. This phenomenon is identical to the delay in the production of light nuclei we saw above, replayed on the atomic scale. Eventually, however, the number of photons with energy above $\epsilon_0$ has redshifted sufficiently for neutral hydrogen to form. This epoch is known as *recombination*, although that is clearly a misnomer: electrons and protons combine, but not re-combine since this is the first time neutral atoms form in the universe.

Helium is another asterisk to the qualitative account given above and the more quantitative one we turn to below. Helium atoms capture single electrons much earlier than do hydrogen atoms since the binding energy is $Z^2\epsilon_0 = 54.4$ eV. The binding energy of the second electron—the one that makes helium neutral—at 24 eV is also larger than 13.6 eV, so the full recombination to neutral helium occurs earlier than that of neutral hydrogen. However, since there are relatively few helium atoms, the vast majority of electrons remain free. For the purposes of decoupling, then, helium recombination plays only a small role, so we will neglect helium in what follows. Percent-level predictions for the CMB anisotropies do need to account for helium.

As long as the reaction[2] $e^- + p \leftrightarrow \mathrm{H} + \gamma$ remains in equilibrium, the condition in Eq. (4.9) (with $1 = e$, $2 = p$, $3 = \mathrm{H}$) ensures that

$$\frac{n_e n_p}{n_\mathrm{H}} = \frac{n_e^{(0)} n_p^{(0)}}{n_\mathrm{H}^{(0)}}. \tag{4.32}$$

This is the *Saha equation*. We can go further here by recognizing that the neutrality of the universe ensures that $n_e = n_p$. Let us define the free electron fraction

$$X_e \equiv \frac{n_e}{n_e + n_\mathrm{H}} = \frac{n_p}{n_p + n_\mathrm{H}}, \tag{4.33}$$

where the denominator is equal to the total number of protons (again neglecting helium). Using Eq. (4.5) for the quantities on the right-hand side of Eq. (4.32) leads to

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_e + n_\mathrm{H}} \left[ \left( \frac{m_e T}{2\pi} \right)^{3/2} e^{-[m_e + m_p - m_\mathrm{H}]/T} \right] \tag{4.34}$$

where we have made the familiar approximation of neglecting the small mass difference of H and $p$ in the prefactor. The argument of the exponential is $-\epsilon_0/T$. The denominator $n_e + n_\mathrm{H}$ (or $n_p + n_\mathrm{H}$) is equal to the baryon density, $\eta_\mathrm{b} n_\gamma \sim 10^{-9} T^3$. So when the temperature is of order $\epsilon_0$, the right-hand side is of order $10^9 (m_e/T)^{3/2} \simeq 10^{15}$. In that case, Eq. (4.34) can be satisfied only if the denominator on the left is very small, that is if $X_e$ is very close to 1: all hydrogen is ionized. Only when the temperature drops far below $\epsilon_0$ does appreciable recombination take place. As $X_e$ falls, the rate for recombination also falls, so that equilibrium becomes more difficult to maintain. Thus, in order to follow the free electron fraction

[2] In the following, $p$ stands for free protons and H for neutral hydrogen, i.e., a proton with an electron attached.

accurately, we need to solve the Boltzmann equation, just as we did for the neutron–proton ratio.

In this case, Eq. (4.8) for the electron density becomes

$$\begin{aligned} a^{-3}\frac{d(n_e a^3)}{dt} &= n_e^{(0)} n_p^{(0)} \langle \sigma v \rangle \left\{ \frac{n_{\rm H}}{n_{\rm H}^{(0)}} - \frac{n_e^2}{n_e^{(0)} n_p^{(0)}} \right\} \\ &= n_{\rm b} \langle \sigma v \rangle \left\{ (1 - X_e) \left( \frac{m_e T}{2\pi} \right)^{3/2} e^{-\epsilon_0/T} - X_e^2 n_{\rm b} \right\} \end{aligned} \tag{4.35}$$

where the last line follows since the ratio $n_e^{(0)} n_p^{(0)} / n_{\rm H}^{(0)}$ is equal to the term in square brackets in Eq. (4.34). Meanwhile, since $n_{\rm b} a^3$ is constant it can be passed through the derivative on the left after expressing $n_e$ as $n_{\rm b} X_e$, so that

$$\frac{dX_e}{dt} = \left\{ (1 - X_e)\beta - X_e^2 n_{\rm b} \alpha^{(2)} \right\} \tag{4.36}$$

where the ionization rate is typically denoted

$$\beta \equiv \langle \sigma v \rangle \left( \frac{m_e T}{2\pi} \right)^{3/2} e^{-\epsilon_0/T} \tag{4.37}$$

and the recombination rate

$$\alpha^{(2)} \equiv \langle \sigma v \rangle . \tag{4.38}$$

The recombination rate has superscript $^{(2)}$ because recombination to the ground state ($n = 1$) is not relevant. Ground-state recombinations lead to production of an ionizing photon, and this photon immediately ionizes a neutral atom. The net effect of such a recombination is zero: no new neutral atoms are formed this way. The only way for recombination to proceed is via capture to one of the excited states of hydrogen; to a good approximation (see Exercise 4.7), this rate is

$$\alpha^{(2)} = 9.78 \frac{\alpha^2}{m_e^2} \left( \frac{\epsilon_0}{T} \right)^{1/2} \ln \left( \frac{\epsilon_0}{T} \right) . \tag{4.39}$$

The Saha approximation, Eq. (4.34), does a good job predicting the redshift of recombination, but fails as the electron fraction drops and the system goes out of equilibrium. Therefore, the detailed evolution of $X_e$ must be obtained by a numerical integration of Eq. (4.36) (Exercise 4.7). Results from a numerical integration including additional complications in the recombination rate (see Exercise 4.7) as well as helium are shown in Fig. 4.4.

We have seen that the neutron/proton ratio affects the abundance of light elements today. Similarly, the evolution of the free electron abundance has major ramifications for observational cosmology. Recombination at $z_* \sim 1000$ is tied to the decoupling of photons

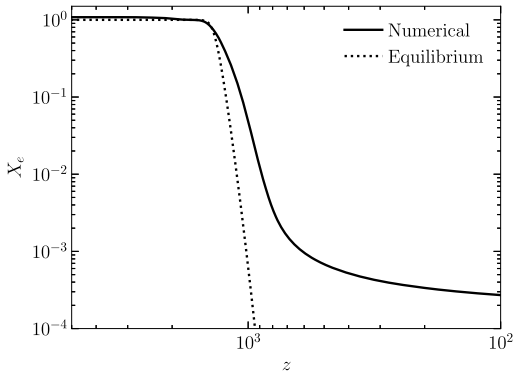



**FIGURE 4.4** Free electron fraction as a function of redshift. The solid line shows the full numerical solution in the fiducial cosmology (given by the CLASS code), while the dotted line is the equilibrium result (the Saha approximation, Eq. (4.34)). Recombination takes place at $z \sim 1000$ corresponding to $T \simeq 0.23$ eV. The equilibrium result correctly identifies the redshift of recombination, but not the evolution of $X_e$, which is important for an accurate prediction of the CMB anisotropies.

from matter.[3] This decoupling, in turn, directly affects the pattern of anisotropies in the CMB that we observe today. Nowadays, sophisticated calculations of the time evolution of $X_e(t)$ are employed in the numerical calculation of CMB anisotropies, which are accurate to the subpercent level.

Now that we have a clear understanding of the evolution of $X_e$, we can move on to determine the epoch of decoupling. Decoupling occurs roughly when the rate for photons to Compton scatter off electrons becomes smaller than the expansion rate.[4] The scattering rate is

$$n_e \sigma_{\rm T} = X_e n_{\rm b} \sigma_{\rm T} \tag{4.40}$$

where $\sigma_{\rm T} = 0.665 \times 10^{-24}$ cm$^2$ is the Thomson cross section, and we continue to ignore helium, thereby assuming that the total number of hydrogen nuclei (free protons + hydrogen atoms) is equal to the total baryon number. Since the ratio of the baryon density to the critical density is $m_p n_{\rm b}/\rho_{\rm cr} = \Omega_{\rm b} a^{-3}$, $n_{\rm b}$ can be eliminated in Eq. (4.40) in favor of $\Omega_{\rm b}$:

$$n_e \sigma_{\rm T} = 7.477 \times 10^{-30}\ {\rm cm}^{-1}\, X_e\, \Omega_{\rm b} h^2\, a^{-3}. \tag{4.41}$$

Dividing by the expansion rate leads to

$$\frac{n_e \sigma_{\rm T}}{H} = \frac{n_e \sigma_{\rm T}}{H_0}\,\frac{H_0}{H} = 0.0692\, a^{-3} X_e\, \Omega_{\rm b} h \frac{H_0}{H}. \tag{4.42}$$

[3] Notice from Fig. 1.4 that even though photons stop scattering off electrons at $z \sim 1000$, electrons do scatter many times off photons until much later. This is not a contradiction: there are many more photons than electrons.

[4] In Ch. 9 we will define a more precise measure of decoupling, making use of the *visibility function*, the probability that a photon last scattered at a given redshift. Using the visibility function, we will show that a CMB photon today most likely last scattered at a slightly higher redshift than inferred by the simple estimate made here.

The ratio on the right depends on the Hubble rate, which is given in Eq. (1.3). From that equation or from Fig. 1.3, we see that at early times, the main contribution comes from either matter or radiation, so $H/H_0 = \Omega_{\rm m}^{1/2} a^{-3/2}[1 + a_{\rm eq}/a]^{1/2}$. Therefore,

$$\frac{n_e \sigma_{\rm T}}{H} = 123 X_e \left(\frac{\Omega_{\rm b} h^2}{0.022}\right) \left(\frac{0.14}{\Omega_{\rm m} h^2}\right)^{1/2} \left(\frac{1+z}{1000}\right)^{3/2} \left[1 + \frac{1+z}{3360}\frac{0.14}{\Omega_{\rm m} h^2}\right]^{-1/2}. \tag{4.43}$$

At large redshift $z \gg 10^3$, $X_e = 1$ so the scattering rate is much larger than the expansion rate. As $z$ drops to $10^3$, $X_e$ begins to drop precipitously, and Eq. (4.43) shows that the scattering rate drops below the expansion rate when $X_e$ drops below $\sim 10^{-2}$; this is the epoch when photons decouple. From Fig. 4.4, we see that $X_e$ drops very quickly from unity to $10^{-3}$. Therefore, decoupling takes place during recombination.

Let us forget all we just learned and ask what would happen if the universe remained ionized throughout its history. In that hypothetical case, $X_e = 1$, and Eq. (4.43) can be trivially solved to find the redshift of decoupling. Setting the right-hand side to 1 leads to

$$1 + z_{\rm decouple} = 39 \left(\frac{0.022}{\Omega_{\rm b} h^2}\right)^{2/3} \left(\frac{\Omega_{\rm m} h^2}{0.14}\right)^{1/3} \quad \text{(no recombination)}. \tag{4.44}$$

Eq. (4.44) tells us that even if the gas had remained ionized throughout the history of the universe, eventually the photons would have decoupled simply because expansion made it more difficult to find the increasingly dilute electrons.

The bulk of the diffuse gas in the universe today is ionized. So, at some point in the universe's history, *reionization* of hydrogen must have taken place. Observations of the most distant quasars suggest that reionization took place at $z > 6$ (Bouwens et al., 2015). We will see in Ch. 9 and Ch. 10 that the Compton scattering of photons re-enabled after reionization leads to imprints in the CMB which can be used to constrain when reionization happened, with the best current measurements (Planck Collaboration, 2018b) pointing to reionization at $z < 10$. The details of this last phase transition of the universe are still very much an open question, however.

## 4.4 Dark matter

We saw in Ch. 1 that there is strong evidence for non-baryonic dark matter in the universe, with $\Omega_{\rm c} \simeq 0.26$. The evidence is compelling but purely gravitational so that it gives very few clues about the identity of the dark matter. This fact, coupled with the creativity of theorists, has led to proposed dark matter candidates spanning some 90 orders of magnitude in mass! There are a corresponding slew of experiments devoted to searching for particles (or other entities) that could be the dark matter. It is impossible to cover the physical motivation behind all of these candidates, so in the bulk of this section we will focus on just one: a weakly interacting massive particle (WIMP). For decades, the WIMP was the odds-on favorite to be the dark matter, and that may still hold true, but the lack of evidence for new

particles in the relevant mass range has reduced the odds considerably and sent the community searching for other candidates. Nonetheless, an understanding of why the WIMP attained its favored status is easy to obtain using the tools we developed in this chapter.

The story of a generic WIMP "$X$" begins at very early times (high temperatures), when $X$ was in equilibrium with the rest of the cosmic plasma, but then experienced *freeze-out* as the reaction rate for annihilation dropped below the expansion rate. Indeed, were the particle kept in equilibrium indefinitely, its abundance would be suppressed by $e^{-m_X/T}$: there would be no $X$ particles in the observable universe. The purpose of this section, then, is to solve the Boltzmann equation for such a particle, determining the epoch of freeze-out and its relic abundance. The hope is that, by fixing its relic abundance so that $\Omega_{\rm X} \simeq 0.26$, we will learn something about the fundamental properties of the particle, such as its mass $m_X$ and annihilation cross section. We then might use this knowledge to detect the particles in a laboratory.

In the generic WIMP scenario, two heavy particles $X$ can annihilate producing two light (essentially massless) particles $\psi$ that are part of the Standard Model (e.g., they could be photons or neutrinos or quarks). The light particles are assumed to be very tightly coupled to the cosmic plasma, so they are in complete equilibrium (chemical as well as kinetic), with $n_\psi = n_\psi^{(0)}$. There is then only one unknown, $n_X$, the abundance of the heavy particle. We can use Eq. (4.8) to solve for this abundance:

$$a^{-3}\frac{d(n_X a^3)}{dt} = \langle\sigma v\rangle\left\{\left(n_X^{(0)}\right)^2 - (n_X)^2\right\}. \tag{4.45}$$

To go further, recall that the temperature typically scales as $a^{-1}$, so if we multiply and divide the factor of $n_X a^3$ inside the parentheses on the left by $T^3$, we can remove $(aT)^3$ outside the derivative, leaving $T^3 d(n_X/T^3)/dt$. Let us then define

$$Y \equiv \frac{n_X}{T^3}. \tag{4.46}$$

The differential equation for $Y$ becomes

$$\frac{dY}{dt} = T^3\langle\sigma v\rangle\{Y_{\rm EQ}^2 - Y^2\}, \tag{4.47}$$

with $Y_{\rm EQ} \equiv n_X^{(0)}/T^3$. As in Sect. 4.2, it is convenient to introduce a new time variable,

$$x \equiv m_X/T \tag{4.48}$$

since $m_X$ sets a rough scale for the temperature during the epoch of interest. Very high temperature corresponds to $x \ll 1$, in which case reactions proceed rapidly so $Y \simeq Y_{\rm EQ}$. Since the $X$ particles are relativistic at these epochs, the $m \ll T$ limit of Eq. (4.5) implies that $Y \simeq 1$. For high $x$, the equilibrium abundance $Y_{\rm EQ}$ becomes exponentially suppressed ($e^{-x}$). Ultimately, $X$ particles will become rare because there are not enough Standard Model particles of sufficient energy to produce a pair of $X$ particles, which requires a

center-of-mass energy of $2m_X$. To change from $t$ to $x$, we need the Jacobian $dx/dt = Hx$. For WIMPs with masses of GeV or larger, dark matter production occurs deep in the radiation era where the energy density scales as $T^4$, so $H = H(m_X)/x^2$. Then the evolution equation becomes

$$\frac{dY}{dx} = -\frac{\lambda}{x^2}\left\{Y^2 - Y_{\rm EQ}^2\right\}, \tag{4.49}$$

where the ratio of the annihilation rate to the expansion rate is parameterized by

$$\lambda \equiv \frac{m_X^3\langle\sigma v\rangle}{H(m_X)}. \tag{4.50}$$

In many theories $\lambda$ is a constant, but in some, the thermally averaged cross section is temperature dependent; this leads to slight numerical changes in the following but unchanged qualitative solutions.

Eq. (4.49) is a form of the Riccati equation, for which in general there are no analytic solutions. In this case, though, we can make use of our understanding of the freeze-out process to get an analytic expression for the final freeze-out abundance $Y_\infty \equiv Y(x=\infty)$. Let us review this understanding in the context of Eq. (4.49). For $x \sim 1$, the left-hand side is of order $Y$ while the right is of order $Y^2\lambda$. We will see that $\lambda$ is typically quite large, so as long as $Y$ is not too small, the right-hand side must zero itself by setting $Y = Y_{\rm EQ}$. At late times, as $Y_{\rm EQ}$ drops precipitously, the terms on the right-hand side will no longer be much larger than the one on the left. In fact, well after freeze-out, $Y$ will be much larger than $Y_{\rm EQ}$: the $X$ particles will not be able to annihilate fast enough to maintain equilibrium. Thus at late times,

$$\frac{dY}{dx} \simeq -\frac{\lambda Y^2}{x^2} \quad (x \gg 1). \tag{4.51}$$

We integrate this from the epoch of freeze-out $x_f$ until very late times $x = \infty$ to get

$$\frac{1}{Y_\infty} - \frac{1}{Y_f} = \frac{\lambda}{x_f}. \tag{4.52}$$

Typically, $Y$ at freeze-out, denoted as $Y_f$, is significantly larger than $Y_\infty$, so a simple approximation is

$$Y_\infty \simeq \frac{x_f}{\lambda}. \tag{4.53}$$

This yields the scaled abundance of $X$ in terms of the freeze-out temperature, which we have not determined. Although more precise determinations are possible (Exercise 4.8), a simple order-of-magnitude estimate for the dark matter problem is $x_f \simeq 20$.

Fig. 4.5 shows the numerical solution to Eq. (4.49) for two different values of $\lambda$. The abundances do track the equilibrium abundances until $m_X/T \sim 10$, after which they level off to a constant. The rough estimate $Y_\infty \sim x_f/\lambda$ is seen to be a reasonable approximation for the relic abundance. Note that particles with larger cross sections (e.g. in the figure $\lambda = 10^9$) freeze out later, and this later freeze-out implies a lower relic abundance.

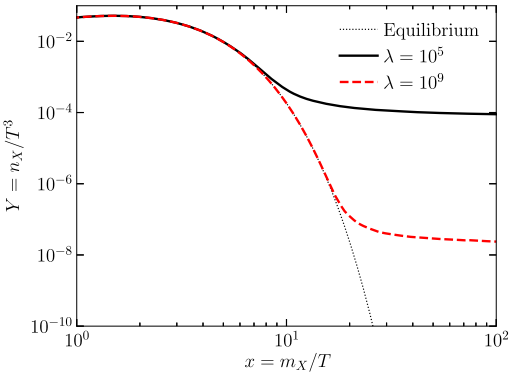



**FIGURE 4.5** Abundance of a heavy stable particle $X$ as a function of inverse temperature $m_X/T$. The dotted line shows the equilibrium abundance. Two different solid curves show the heavy particle abundance for two different values of $\lambda$, the ratio of the annihilation rate to the Hubble rate.

There is one more piece of physics needed in order to determine the present-day abundance of these heavy particle relics. After freeze-out, the heavy particle density simply falls off as $a^{-3}$. So its energy density today (at $a_0 = 1$) is equal to $m_X(a_1/a_0)^3$ times its number density at $a_1$, where $a_1$ corresponds to a time sufficiently late that $Y$ has reached its asymptotic value, $Y_\infty$. The number density at that time is $Y_\infty T_1^3$, so

$$\rho_{X,0} = m_X Y_\infty T_0^3 \left(\frac{a_1 T_1}{a_0 T_0}\right)^3 \simeq \frac{m_X Y_\infty T_0^3}{30}. \tag{4.54}$$

The second equality here is nontrivial. You might expect that $aT$ remains constant through the evolution of the universe, so that the ratio $a_1 T_1/a_0 T_0$ would be unity. It is not, for the same reason that the CMB and neutrinos have different temperatures. We saw in Ch. 2 that photons are heated by $e^\pm$ annihilation, while neutrinos which have already decoupled are not. Similarly, as the universe expands, photons are heated by the annihilation of the zoo of particles with masses between 1 MeV and $m_X$, which we will assume to be larger than $\sim$100 GeV, so $T$ does not fall simply as $a^{-1}$. You can show in Exercise 4.9 that as a result $(a_1 T_1/a_0 T_0)^3 \simeq 1/30$. Finally, to find the fraction of critical density today contributed by $X$, insert our expression for $Y_\infty$ and divide by $\rho_{\rm cr}$:

$$\begin{aligned}\Omega_X &= \frac{x_f}{\lambda}\frac{m_X T_0^3}{30\rho_{\rm cr}} \\ &= \frac{H(m_X) x_f T_0^3}{30 m_X^2 \langle\sigma v\rangle \rho_{\rm cr}}.\end{aligned} \tag{4.55}$$

To find the present density of heavy particles, then, we need to compute the Hubble rate when the temperature was equal to the $X$ mass, $H(m_X)$, for which we need the energy density when the temperature was equal to $m_X$. The energy density in the radiation era is

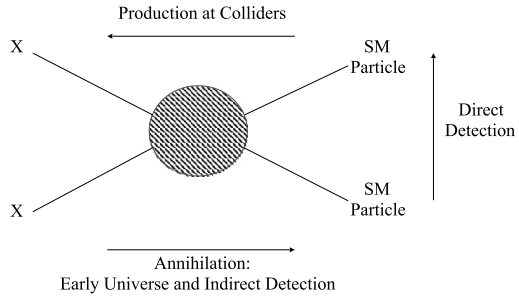



**FIGURE 4.6** WIMPs ($X$) can annihilate into Standard Model (SM) particles in the early universe leading to a calculable relic abundance, as discussed in this section. The same type of process enables *indirect detection*, where we observe the products of the annihilation of two dark matter particles in space. Annihilations are mediated by some unknown fundamental interactions (depicted by the hatched region in the center). These also enable the reverse process, where two Standard Model particles produce two WIMPs; searches for these are ongoing at colliders. Finally, a third search method is to turn the interaction on its side and search for the recoil of a nucleus that has been hit by a WIMP, so-called *direct detection*.

given by Eq. (4.25) with $g_*$ a function of the temperature. Therefore,

$$\Omega_X = \left[\frac{4\pi^3 G g_*(m_X)}{45}\right]^{1/2} \frac{x_f T_0^3}{30\langle\sigma v\rangle \rho_{\rm cr}}. \tag{4.56}$$

We see that $\Omega_X$ does not explicitly depend on the mass of the $X$ particle.[5] So it is mainly the cross section that determines the relic abundance.

Let us now see what order of magnitude is needed to get dark matter today, i.e., to get $\Omega_X h^2 \simeq 0.1$. At the temperatures of interest for dark matter production, $T \sim 100$ GeV, $g_*(m_X)$ includes contributions from all the particles in the Standard Model (three generations of quarks and leptons, photons, gluons, weak bosons, and the Higgs boson) and so is of order 100. Normalizing $g_*(m_X)$ and $x_f$ by their nominal values leads to

$$\Omega_X h^2 = 0.1 \left(\frac{x_f}{20}\right) \left(\frac{g_*(m_X)}{100}\right)^{1/2} \frac{2\times 10^{-26}\ {\rm cm}^3\,{\rm sec}^{-1}}{\langle\sigma v\rangle} \tag{4.57}$$

where we have reinstated the speed of light to get the dimensions of $\langle\sigma v\rangle$ correct. The fact that this estimate gives the correct relic abundance for $\langle\sigma v\rangle \sim 10^{-26}\ {\rm cm}^3/\ {\rm s}$ is taken as a good sign: there are several theories that predict the existence of particles with cross sections in this range.

Apart from the abundance of dark matter, a further important property we can calculate is the temperature of the dark matter today. In Exercise 4.10, you will find that it is very small: WIMPs are an example of *cold dark matter*, which is what is required to explain the structure in the universe we see today.

[5] There is a weak implicit dependence on mass in the freeze-out temperature $x_f$ and in $g_*$, which is to be evaluated when $T = m_X$.

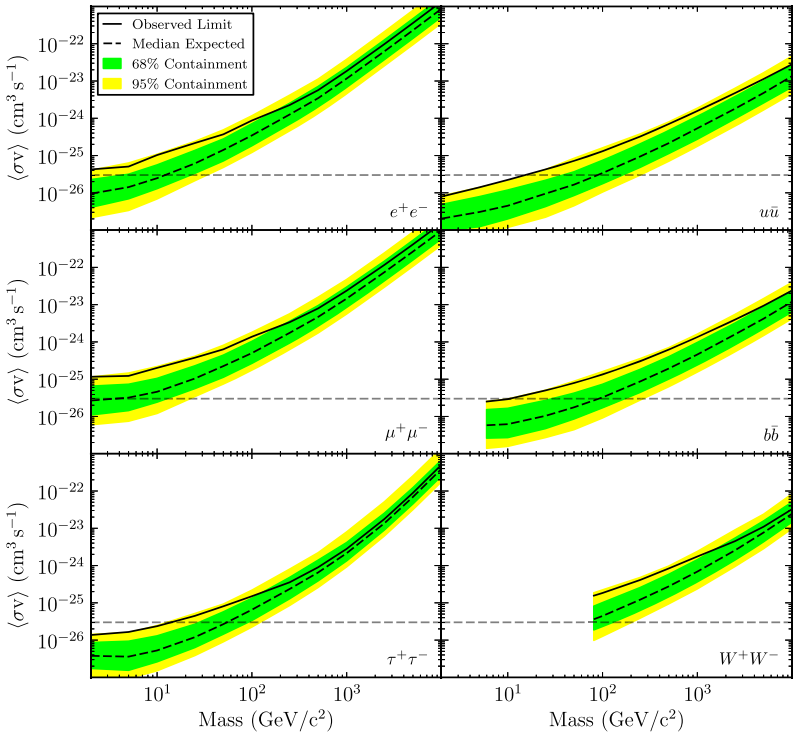



**FIGURE 4.7** Constraints on the dark matter annihilation cross section from Fermi $\gamma$-ray observations as a function of mass for different annihilation channels (the final states are indicated in each panel). Regions above the solid curve are excluded, and the horizontal dashed line gives the target, the cross section needed to produce the correct relic abundance (Eq. (4.57)). From Ackermann et al. (2014).

The very interaction—annihilation into Standard Model particles—that determines the relic density of a WIMP also opens up three possibilities for detection, as depicted in Fig. 4.6. The first is to exploit the annihilation process itself. Two dark matter particles in space can annihilate and produce Standard Model particles that will leave a signature in a detector. The exact nature of the signature depends on the details of the model; for example, dark matter particles that annihilate into quarks and anti-quarks will ultimately produce a shower of lighter particles including photons. Telescopes such as the Fermi Large Area Telescope are sensitive to high-energy ($\gamma$-ray) photons so can detect this signal, which has been dubbed *indirect detection* (in contrast to the more "direct" measurement discussed below). Fig. 4.7 shows the constraints from Fermi data for 6 different annihilation channels. These were obtained by focusing on small nearby galaxies that have relatively large dark matter abundances (dwarf galaxies), as inferred from the velocity dispersions of the stars. Note that indirect detection is more powerful in excluding low mass dark matter particles (since the interaction rate scales as $n_X^2 \propto 1/m_X^2$ at fixed dark matter density), and indeed for many channels has excluded regions below $m_X \sim 10$ GeV.

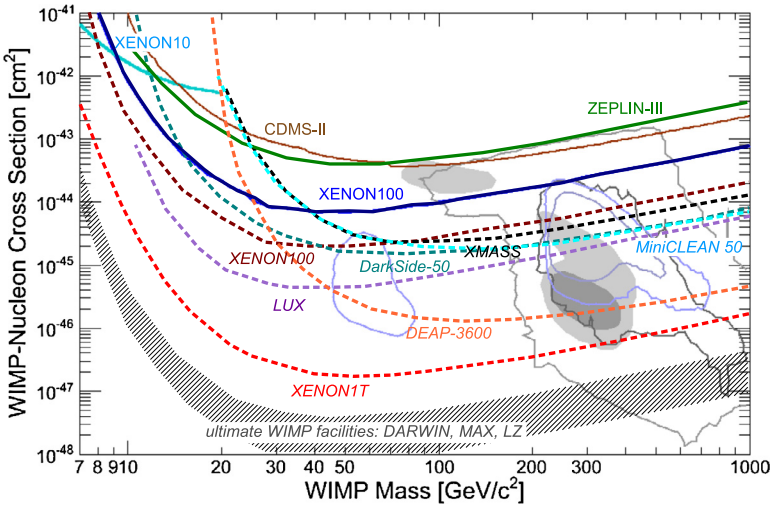



**FIGURE 4.8** Constraints on the dark matter-nucleus scattering cross section from various low-background laboratory experiments (cross sections above the solid lines are excluded), as a function of WIMP mass. Contours show the expected parameter space in a subset of *supersymmetry* models that have natural WIMP candidates (Buchmueller et al., 2012). The dashed lines were projections at the time this plot was produced, but the constraints have since been largely attained, indicating the rapid progress in the field. From Schumann (2012).

Fig. 4.6 shows that two other detection possibilities are available: for one, the reverse process in colliders, where colliding high-energy protons could produce pairs of massive dark matter particles. Finally, we can flip the interaction on its edge, where an incoming dark matter particle in our Galaxy imparts some of its kinetic energy to a nucleus in a large detector. This process is exploited in *direct detection* experiments. Although it is challenging to detect these reactions because the rate is so low, there has been enormous progress over the past several decades by introducing novel techniques for separating the signal from the backgrounds, by building larger detectors, and by housing these in underground sites to further reduce backgrounds. Fig. 4.8 gives an indication of this progress in several ways. First, a similar plot shown in the first edition of this book (published about ten years before Fig. 4.8 was made) showed constraints that limited the cross section[6] to be above $10^{-41}$ cm$^2$. In just ten years then, the constraints improved by almost three orders of magnitude. The dashed contours give a sense of where Schumann (2012) projected the constraints would lie when experiments such as LUX and XENON ramped up. And, indeed in 2019, the experiments had approached those expected marks, another three orders of magnitude tighter. The hatched region at the bottom of the plot has not yet been attained, but likely will be. Beneath that, the background of cosmic neutrinos provides a natural ex-

[6] The dark matter–nucleon cross section that is constrained by direct detection experiments differs from the annihilation cross section that determines the relic abundance, although in a given model, both can be calculated.

perimental limit on how low we will be able to probe WIMP cross sections through direct detection.

Given how rapidly WIMP models are being eliminated by direct detection experiments and the lack of evidence for new heavy particles at colliders, the community is investigating a wide variety of alternatives to WIMP dark matter. The thermal freeze-out scenario puts tight constraints on the mass and cross section of the dark matter particle. However, there are also nonthermal means of producing dark matter. One example is the axion, a fundamental particle hypothesized for independent reasons, that—if it exists—was produced during the phase transition where quarks condensed into baryons. Axions, as bosons, are produced in the ground state, forming a Bose–Einstein condensate. One of the interesting features of axions is that they can serve as cold dark matter even though they are very light ($\ll 1$ eV), and experimental efforts are under way to look for evidence of axion dark matter. In the meantime, cosmologists have begun to investigate dark matter candidates as light as $10^{-21}$ eV, and as heavy as several solar masses (in the form of black holes)! The perceived troubles of the WIMP have opened up a slew of possibilities for new dark matter candidates and innovative ways of detecting them.

## 4.5 Summary

The light elements in the universe formed when the temperature of the cosmic plasma was of order 0.1 MeV. Since the completion of BBN, roughly a quarter of the mass of the baryons is in the form of $^4$He, the remaining in the form of free protons with only trace amounts of deuterium, $^3$He, and lithium.

The universe remained ionized until the temperature dropped well below the ionization energy of hydrogen. The epoch of recombination—at which time electrons and protons combined to form neutral hydrogen—is at redshift $z_* \sim 1100$ corresponding to a temperature $T_* \sim 0.25$ eV. Before recombination, photons and electrons and protons were tightly coupled with one another because of Compton and Coulomb scattering. After this time, photons traveled freely through the universe without interacting, so the photons in the CMB we observe today offer an excellent snapshot of the universe at $z_*$.

The details of both nucleosynthesis and recombination are heavily influenced by the fact that the reactions involved eventually become too slow to keep up with the expansion rate. This feature may also be responsible for the production of dark matter in the universe. We explored the popular scenario wherein a massive, neutral stable particle stops annihilating when the temperature drops significantly beneath its mass. The present-day abundance of such a particle can be determined in terms of its annihilation cross section, as in Eq. (4.57). Larger cross sections correspond to more efficient annihilation and therefore a lower abundance today. Roughly, thermally averaged cross sections $\langle \sigma v \rangle \sim 2 \times 10^{-26}$ cm$^3$ sec$^{-1}$ are needed to match the dark matter abundance observed today. Such cross sections and the requisite stable, neutral particles emerge fairly naturally in extensions of the Standard Model of particle physics, such as supersymmetry, although no (non-gravitational) evidence of them has so far been found in the laboratory, in as-

trophysical objects, or in accelerators (Figs. 4.7–4.8). One thus should not discount other well-motivated scenarios where dark matter is produced non-thermally, such as axions.

Despite the existence of a large spectrum of viable dark matter models, the predictions for cosmological observables we will derive in the following chapters are entirely independent of the microscopic nature of dark matter. They only rely on two properties of dark matter: that it is cold and weakly interacting. Apart from these properties, all predictions only depend on a single, well-constrained number, the overall amount of dark matter parametrized by $\Omega_{\rm c}$.

## Exercises

**4.1** Compute the equilibrium number density (i.e., with zero chemical potential) of a species with mass $m$ and degeneracy $g = 2$ in the limits of large and small $m/T$ for both bosons and fermions. You will find Eqs. (C.29) and (C.30) helpful for the high-$T$ Bose–Einstein and Fermi–Dirac limits.

**4.2** Track the $e^{\pm}$ density through annihilation assuming $n_{e^\pm} = n^{(0)}_{e^\pm}$. This assumption holds during the BBN epoch because electromagnetic interactions (e.g., $e^+ + e^- \leftrightarrow \gamma + \gamma$) keep them in equilibrium. When does the density fall to 1% of the photon energy density? If $\eta_{\rm b} \simeq 6 \times 10^{-10}$, at what temperature do you expect $n_{e^-}$ to depart from $n^{(0)}_{e^-}$?

**4.3** Suppose that there were no baryon asymmetry so that the number density of baryons exactly equaled that of anti-baryons. Determine the final relic density of (baryons+anti-baryons). At what temperature is this asymptotic value reached?

**4.4** Compute the rate for neutron-to-proton conversion, $\lambda_{np}$, following the steps given below. There are two processes which contribute to $\lambda_{np}$: $n + \nu_e \to p + e^-$ and $n + e^+ \to p + \overline{\nu}_e$. Assume that all particles can be described by Boltzmann statistics and neglect the mass of the electron. With these approximations the two rates are identical.

**(a)** Use Eq. (4.7) to write down the rate for $n + \nu_e \to p + e^-$. Perform the integrals over heavy particle momenta to get

$$\lambda_{np} = n^{(0)}_{\nu_e} \langle \sigma v \rangle = \frac{\pi}{4m^2} \int \frac{d^3 p_\nu}{(2\pi)^3 2 p_\nu} e^{-p_\nu/T} \times \int \frac{d^3 p_e}{(2\pi)^3 2 p_e} \delta^{(1)}_{\rm D}(\mathcal{Q} + p_\nu - p_e) |\mathcal{M}|^2 . \tag{4.58}$$

**(b)** The amplitude squared is equal to $|\mathcal{M}|^2 = 32 G_{\rm F}^2 (1 + 3 g_A^2) m_p^2 p_\nu p_e$, where $g_A$ is the axial-vector coupling of the nucleon. This can be related to the neutron lifetime via $\tau_n^{-1} = \lambda_0 G_{\rm F}^2 (1 + 3 g_A^2) m_e^5 / (2\pi^3)$, where the phase-space integral

$$\lambda_0 \equiv \int_1^{\mathcal{Q}/m_e} dx\, x (x - \mathcal{Q}/m_e)^2 (x^2 - 1)^{1/2} = 1.636. \tag{4.59}$$

Carry out the integrals in Eq. (4.58) to get the rate, $\lambda_{np}$ in terms of $\tau_n$. Do not forget to multiply by 2 for the two different reactions. You should obtain (Bernstein, 2004)

$$\lambda_{np} = \frac{255}{\tau_n x^5}(12 + 6x + x^2). \tag{4.60}$$

**4.5** Solve the rate equation (4.26) numerically to determine the neutron fraction as a function of the temperature. Ignore decays. For this, use the rate $\lambda_{np}$ derived in the previous exercise, Eq. (4.60). Plug in numbers to show that the Hubble rate at $x = 1$ is

$$H(x = 1) = \sqrt{\frac{4\pi^3 G Q^4}{45}} \times \sqrt{10.75} = 1.13\ \mathrm{s}^{-1}. \tag{4.61}$$

Solve the ordinary differential Eq. (4.26) numerically. Alternatively you can follow the semi-analytic route given by Bernstein et al. (1989). Compare your results (of either approach) with Fig. 4.3, and the asymptotic result at $x = \infty$ with the result in the text, $X_n(x = \infty) = 0.15$.

**4.6** Determine $\eta_\mathrm{b}$ in terms of $\Omega_\mathrm{b}h^2$. Show that it is given by Eq. (4.10).

**4.7** Solve for the evolution of the free electron fraction during recombination. Do not compare your results with Fig. 4.4 until you finish part **(d)**. Assume the fiducial Euclidean ΛCDM cosmology.

**(a)** Use as an evolution variable $x \equiv \epsilon_0/T$ instead of time in Eq. (4.36). Rewrite the equation in terms of $x$ and the Hubble rate at $T = \epsilon_0$.

**(b)** Using the methods of Sect. 4.4, find the final freeze-out abundance of the free electron fraction, $X_e(x = \infty)$.

**(c)** Numerically integrate the equation from **(a)** from $x = 1$ down to $x = 1000$. What is the final frozen-out $X_e$?

**(d)** Peebles (1968) argued that even captures to excited states of the hydrogen atom would not be important except for the fraction of times that the $n = 2$ state decays into two photons, or expansion redshifts the Lyman alpha photon ($n = 2 \to 1$) so that it cannot pump up a ground-state atom. Quantitatively, he multiplied the right-hand side of Eq. (4.36) by the correction factor,

$$C = \frac{\Lambda_\alpha + \Lambda_{2\gamma}}{\Lambda_\alpha + \Lambda_{2\gamma} + \beta^{(2)}} \tag{4.62}$$

where the two-photon decay rate is $\Lambda_{2\gamma} = 8.227\ \mathrm{s}^{-1}$; Lyman alpha production is $\beta^{(2)} = \beta e^{3\epsilon_0/4T}$; and

$$\Lambda_\alpha = \frac{H(3\epsilon_0)^3}{n_\mathrm{H}(8\pi)^2} \tag{4.63}$$

where $H$ is the expansion rate and $n_\mathrm{H}$ is the number density of hydrogen (which you can set to $n_\mathrm{b}(1 - X_e)$). Do this and show how it changes your final answer.

Now compare the freeze-out abundance with the result of **(c)** and the evolution with Fig. 4.4.

**4.8** Find an approximation to the freeze-out temperature of annihilating heavy particles by setting $x_f$ such that $n^{(0)}(x_f)\langle\sigma v\rangle = H(x_f)$.

**4.9** Typically the temperature of the cosmic plasma cools as $a^{-1}$ with the expansion. However, when particles annihilate, they deposit energy into the plasma, thereby slowing the cooling. Use the fact that the entropy density (Eq. (2.70)) scales as $a^{-3}$ to compute the ratio of $(aT)^3$ at $T = 10$ GeV (a time when WIMPs might have decoupled) to its present value today.

**4.10** In Exercise 3.11, you showed that a thermal distribution of nonrelativistic particles which do not interact has a temperature which scales as $a^{-2}$, as opposed to that of relativistic particles which we have seen scales as $a^{-1}$. So $T_{\rm dm} \propto T^2$. Fix the normalization by requiring $T_{\rm dm} = T$ when each is equal to the dark matter mass. Estimate the typical thermal velocity of a dark matter particle with mass equal to 100 GeV when the photon temperature is 1 eV, and when it is equal to 2.7 K.

# 5

# The inhomogeneous universe: matter & radiation

Starting from this chapter, we will be interested in the anisotropies in the cosmic distribution of photons and inhomogeneities in the matter. We have already become familiar with the equations that we must solve: the Einstein and Boltzmann equations introduced in Ch. 3, with one Boltzmann equation each for each species in the universe. Unlike Ch. 4, wherein we were interested solely in the evolution of the homogeneous number density of the different species, here we must account for the spatial and directional dependence of the distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$. This turns out to complicate the algebra significantly, but, with the tools described in Ch. 3, we are poised to tackle these complications systematically, as essentially one long homework problem. The set of equations we will ultimately arrive at is quite simple and of clear physical content.

The photons are affected by gravity and by Compton scattering with free electrons. The electrons are in addition tightly coupled to the protons. Both of these, of course, are also affected by gravity. The metric that determines the gravitational forces is influenced by all these components plus the neutrinos and the dark matter. Thus, to solve for the distributions of any of these components, we need to simultaneously solve for all the other components.

In order not to lose track of the big picture, it is useful to visualize the various interactions described by the Boltzmann and Einstein equations between nucleosynthesis and recombination as in Fig. 5.1. At the end of this chapter, we will have in hand the evolution equations for perturbations in all relevant species in the universe, which takes us a big step closer to calculating actual cosmological observables. The main ingredient missing will be how to solve for the metric perturbations that appear in the Boltzmann equations, that is, gravity, which we will turn to in the next chapter. Hence, this chapter will derive how matter, photons and neutrinos behave in a given expanding spacetime with perturbations.

In principle, we should also include perturbations to the dark energy density, which always exist if the dark energy is *not* a cosmological constant. In practice, though, most viable models of dark energy predict that the perturbations are very small and only became relevant very recently. For our purposes then, we are justified in neglecting the dark energy as a source of perturbations to the metric.

We will begin with the Boltzmann equation for the photons, including a detailed derivation of the collision term. Following a similar pattern, we then derive the Boltzmann equations for dark matter, baryons, and neutrinos.

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00011-5
Copyright © 2021 Elsevier Inc. All rights reserved.

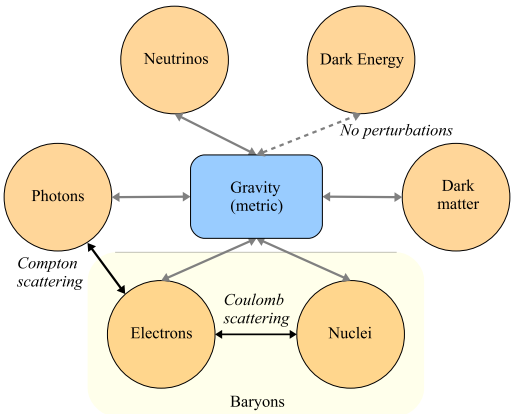



**FIGURE 5.1** The ways in which the different components of the universe interact with each other. The connections are encoded in the coupled Boltzmann–Einstein equations. The tight coupling between electrons and nuclei through Coulomb scattering allows us to treat them as a single component, for which we use the conventional name *baryons*. We do not consider perturbations to the dark energy (which are absent in the case of the cosmological constant), so dark energy only enters in the background metric.

## 5.1 The collisionless Boltzmann equation for photons

We begin with the Boltzmann equation for photons. We have derived the left-hand side of this, at linear order in perturbations, in Sect. 3.3.3, leading to Eq. (3.74):

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - p\frac{\partial f}{\partial p}\left[H + \dot{\Phi} + \frac{\hat{p}^i}{a}\frac{\partial \Psi}{\partial x^i}\right]. \tag{5.1}$$

To go further we must now expand the photon distribution function $f$ about its zeroth-order Bose–Einstein form. We will do this in a way that may seem odd at first. Let us write

$$f(\boldsymbol{x}, p, \hat{\boldsymbol{p}}, t) = \left[\exp\left\{\frac{p}{T(t)[1+\Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, t)]}\right\} - 1\right]^{-1}. \tag{5.2}$$

Here the zeroth-order temperature $T$ is a function of time only, not space. In the smooth universe, photons are distributed homogeneously, so $T$ is independent of $\boldsymbol{x}$, and isotropically, so $T$ is independent of the direction of propagation $\hat{\boldsymbol{p}}$. Now that we want to describe perturbations about this smooth universe, we need to allow for a *perturbation to the distribution function*, which is characterized by the fractional temperature perturbation $\Theta$, which could also be called $\delta T/T$. $\Theta$ allows for inhomogeneities in the photon distribution (it depends on $\boldsymbol{x}$) as well as anisotropies (it also depends on $\hat{\boldsymbol{p}}$). Recall from Sect. 1.5 that in the end we observe the temperature perturbations on our "CMB sky." That is, what we measure is $\Theta$ as a function of $\hat{\boldsymbol{p}}$, which is the arrival direction of the photon, at a fixed location $\boldsymbol{x}_{\rm Earth}$ and time $t_0$: $\delta T/T(\hat{\boldsymbol{p}}) = \Theta(\boldsymbol{x}_{\rm Earth}, \hat{\boldsymbol{p}}, t_0)$.

Note that we assume here that $\Theta$ does *not* depend on the magnitude of the momentum $p$. We will soon see that this is a valid assumption at the order we work in, following

directly from the fact that the magnitude of the photon momentum is virtually unchanged by the dominant form of interaction, Compton scattering.[1]

The perturbation $\Theta$ is small, so in keeping with the assumption of small $\Phi$, $\Psi$ that went into the derivation of Eq. (5.1), we expand Eq. (5.2) to first order in $\Theta$:

$$f(\boldsymbol{x}, \boldsymbol{p}, t) \simeq \frac{1}{e^{p/T(t)} - 1} + \left( \frac{\partial}{\partial T} \left[ \exp\left\{ \frac{p}{T(t)} \right\} - 1 \right]^{-1} \right) T(t) \Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, t)$$
$$= f^{(0)}(p, t) - p \frac{\partial f^{(0)}(p, t)}{\partial p} \Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, t). \tag{5.3}$$

In the last line we have identified the zeroth-order distribution function as the Bose–Einstein distribution with zero chemical potential,

$$f^{(0)} \equiv \left[ \exp\left\{ \frac{p}{T} \right\} - 1 \right]^{-1}, \tag{5.4}$$

and made use of the fact that, for this function, $T \partial f^{(0)}/\partial T = -p \partial f^{(0)}/\partial p$.

We can now separate the Boltzmann equation (5.1) into a zeroth-order equation for $f^{(0)}$, and a first-order equation for the perturbation $\Theta$. The zeroth-order part is nothing but Eq. (3.39) derived in Ch. 3 (without the $\hat{p}^i \partial f/\partial x^i$ term which vanishes at the background level):

$$\left. \frac{df}{dt} \right|_{\text{zero order}} = \frac{\partial f^{(0)}}{\partial t} - Hp \frac{\partial f^{(0)}}{\partial p} = 0. \tag{5.5}$$

We have set $df/dt$ here equal to zero, i.e., set the collision term on the right-hand side of Eq. (3.39) to zero. That corresponds to the statement that the collision terms will be proportional to $\Theta$ and other perturbations to the homogeneous universe. There is a profound reason for this: the zeroth-order distribution function is set precisely by the requirement that the collision term vanishes. Another, perhaps more familiar way of saying this is to point out that any collision term includes the rate for the given reaction and for its inverse. If the distribution functions are set to their equilibrium values, the rate for the reaction precisely cancels the rate for its inverse. If a given component is out of equilibrium, collisions will drive it toward its equilibrium distribution. This is the reason we expected a Bose–Einstein distribution in the first place.

Returning to Eq. (5.5), we can rewrite the time derivative as

$$\frac{\partial f^{(0)}}{\partial t} = \frac{\partial f^{(0)}}{\partial T} \frac{dT}{dt} = -\frac{dT/dt}{T} p \frac{\partial f^{(0)}}{\partial p}$$

[1] Although we will only deal with elastic scattering of photons and electrons here, which is technically known as *Thomson* scattering, we will stick to the more general term "Compton" throughout. We will later encounter inelastic scattering in Sect. 11.3.

so that the zeroth-order equation becomes

$$\left[-\frac{dT/dt}{T}-\frac{\dot{a}}{a}\right]\frac{\partial f^{(0)}}{\partial p}=0. \tag{5.6}$$

Thus $dT/T=-da/a$ or $T\propto 1/a$, a relation that is familiar by now (Sect. 2.4.1).

We now extract the equation for the deviation of the photon temperature from its zeroth-order value, i.e., an equation for $\Theta$, from Eq. (5.1). To do this, we insert the expansion of Eq. (5.3) everywhere we encounter $f$:

$$\begin{aligned}\left.\frac{df}{dt}\right|_{\text{first order}} &= -p\frac{\partial}{\partial t}\left[\frac{\partial f^{(0)}}{\partial p}\Theta\right]-p\frac{\hat{p}^i}{a}\frac{\partial\Theta}{\partial x^i}\frac{\partial f^{(0)}}{\partial p}+H\Theta p\frac{\partial}{\partial p}\left[p\frac{\partial f^{(0)}}{\partial p}\right]\\ &\quad -p\frac{\partial f^{(0)}}{\partial p}\left[\dot{\Phi}+\frac{\hat{p}^i}{a}\frac{\partial\Psi}{\partial x^i}\right].\end{aligned} \tag{5.7}$$

Consider the first term on the right-hand side here. The time derivative on $f^{(0)}$ can be rewritten as a temperature derivative so

$$\begin{aligned}-p\frac{\partial}{\partial t}\left[\frac{\partial f^{(0)}}{\partial p}\Theta\right] &= -p\frac{\partial f^{(0)}}{\partial p}\frac{\partial\Theta}{\partial t}-p\Theta\frac{dT}{dt}\frac{\partial^2 f^{(0)}}{\partial T\partial p}\\ &= -p\frac{\partial f^{(0)}}{\partial p}\frac{\partial\Theta}{\partial t}+p\Theta\frac{dT/dt}{T}\frac{\partial}{\partial p}\left[p\frac{\partial f^{(0)}}{\partial p}\right].\end{aligned} \tag{5.8}$$

The second line follows here since $\partial f^{(0)}/\partial T=-(p/T)\partial f^{(0)}/\partial p$. The second term on this second line cancels the third term on the right in Eq. (5.7), so we can finally write down the left-hand side of the Boltzmann equation for $\Theta$:

$$\left.\frac{df}{dt}\right|_{\text{first order}} = -p\frac{\partial f^{(0)}}{\partial p}\left[\dot{\Theta}+\frac{\hat{p}^i}{a}\frac{\partial\Theta}{\partial x^i}+\dot{\Phi}+\frac{\hat{p}^i}{a}\frac{\partial\Psi}{\partial x^i}\right]. \tag{5.9}$$

The first two terms here correspond to a derivative along light rays (null geodesics) in the homogeneous universe. They describe how the distribution function evolves in the absence of collisions, also known as "free streaming." The last two account for the gravitational effect of perturbations. Note that every time $x$ appears it is multiplied by $a$, the scale factor. This must happen, for physical distances are $ax$. This equation for $\Theta$ is not complete, since we know that at first order in perturbations, there will be a nonzero collision term. We turn to that next.

## 5.2 Collision terms: Compton scattering

Our task in this section is to determine the influence Compton scattering has on the photon distribution function. This follows the general treatment of collision terms in Sect. 3.2.3, and is similar to our applications in Ch. 4, except that we now have to include

perturbations to the distribution functions. Recall that in Ch. 4 we considered processes that are out of *chemical equilibrium*, but could always rely on *kinetic equilibrium*. We will now deal with the absence of kinetic equilibrium as well. This is crucial in order to accurately follow the photon distribution through recombination and hence to the observed CMB anisotropies.

The scattering process of interest is

$$e^-(\boldsymbol{q}) + \gamma(\boldsymbol{p}) \leftrightarrow e^-(\boldsymbol{q}') + \gamma(\boldsymbol{p}'), \tag{5.10}$$

where the momentum of each particle is indicated. We are interested in the photon distribution evaluated at momentum $\boldsymbol{p}$ (with magnitude $p$ and direction $\hat{\boldsymbol{p}}$). Therefore we must integrate over all other momenta $(\boldsymbol{q}, \boldsymbol{q}', \boldsymbol{p}')$ which affect $f(\boldsymbol{p})$, as done in Sect. 3.2.3. From Eq. (3.48), the collision term is

$$\begin{aligned} C[f(\boldsymbol{p})] = \frac{1}{2E(p)} \int \frac{d^3q}{(2\pi)^3 2E_e(q)} \int \frac{d^3q'}{(2\pi)^3 2E_e(q')} \int \frac{d^3p'}{(2\pi)^3 2E(p')} \sum_{3\ \text{spins}} |\mathcal{M}|^2 \\ \times (2\pi)^4 \delta_{\rm D}^{(3)}[\boldsymbol{p} + \boldsymbol{q} - \boldsymbol{p}' - \boldsymbol{q}']\, \delta_{\rm D}^{(1)}\left[E(p) + E_e(q) - E(p') - E_e(q')\right] \\ \times \{f_e(\boldsymbol{q}') f(\boldsymbol{p}') - f_e(\boldsymbol{q}) f(\boldsymbol{p})\}. \end{aligned} \tag{5.11}$$

We have explicitly included the sum over the final spin states of the outgoing electron and the photon (two each) and the electron with which the photon with momentum $\boldsymbol{p}$ scatters. Note that, unlike the case in Sect. 4.1, we have not integrated over the final photon momentum $\boldsymbol{p}$. Again, this reflects our need to understand how photons traveling in different directions interact: we will see that the collision term depends on $\hat{\boldsymbol{p}}$.

In Eq. (5.11), we have neglected stimulated emission and Pauli blocking, which would lead to factors of $1 + f$ (for the photons) and $1 - f_e$ (for the electrons) with the appropriate momenta. Pauli blocking is never important after the time of electron–positron annihilation because the occupation numbers $f_e$ are very small; we have used this fact in Sect. 4.2 as well. We will see below why the stimulated emission factors drop out. The photon energies in Eq. (5.11) are simply $E(p) = p$ and $E(p') = p'$. On the other hand, we assume the nonrelativistic limit for electrons. This is completely sufficient at the time of recombination where the typical kinetic energies of electrons, of order $T$, are much smaller than the electron mass. We thus have

$$\begin{aligned} E(p) = p \sim T\,, \quad \text{while} \\ E_e(q) - m_e = q^2/(2m_e) \sim T \quad \Rightarrow \quad q \sim T\sqrt{\frac{2m_e}{T}}\,. \end{aligned} \tag{5.12}$$

Here, we have used the fact that, close to equilibrium, typical photon energies and kinetic energies of electrons are of order $T$. We see that the electron momenta are much larger than the photon momenta, since $m_e/T \gg 1$.

Now, using the three-dimensional momentum delta function to do the $\boldsymbol{q}'$ integral in Eq. (5.11), we have

$$C[f(\boldsymbol{p})] = \frac{\pi}{2m_e p}\int\frac{d^3q}{(2\pi)^3 2m_e}\int\frac{d^3p'}{(2\pi)^3 2p'}\delta_{\rm D}^{(1)}\left[p + E_e(q) - p' - E_e(|\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}'|)\right] \times \sum_{3\ \text{spins}}|\mathcal{M}|^2\{f_e(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')f(\boldsymbol{p}') - f_e(\boldsymbol{q})f(\boldsymbol{p})\}. \tag{5.13}$$

To go further, we need to understand the kinematics of nonrelativistic Compton scattering. The most important feature of this process for our purposes is that very little energy is transferred. In particular,

$$p' - p = E_e(q) - E_e(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}') = \frac{q^2}{2m_e} - \frac{(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')^2}{2m_e} \simeq \frac{(\boldsymbol{p}'-\boldsymbol{p})\cdot\boldsymbol{q}}{m_e}, \tag{5.14}$$

where the last approximate equality holds since, from Eq. (5.12), $q$ is much larger than $p$ and $p'$. Since $p$ and $p'$ are of the same order, the right-hand side is at most of order $2pq/m_e$ (if $\boldsymbol{p}' \simeq -\boldsymbol{p}$). Using Eq. (5.12), this means that the fractional change in photon energy is at most $|p'-p|/p \lesssim 2q/m_e \sim 2\sqrt{2T/m_e} \ll 1$. Thus, nonrelativistic Compton scattering is nearly elastic and $p' \simeq p$. In the end, this justifies why we have written $\Theta$ as a function of $\hat{\boldsymbol{p}}$ but not $p$. Further, it then makes sense to expand the final electron kinetic energy $(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')^2/(2m_e)$ around its zeroth-order value of $q^2/(2m_e)$. The delta function can be expanded as

$$\begin{aligned}&\delta_{\rm D}^{(1)}\left[p - p' + E_e(q) - E_e(|\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}'|)\right]\\ &\quad\simeq \delta_{\rm D}^{(1)}(p-p') + \frac{(\boldsymbol{p}'-\boldsymbol{p})\cdot\boldsymbol{q}}{m_e}\frac{\partial}{\partial p}\delta_{\rm D}^{(1)}(p-p')\\ &\quad= \delta_{\rm D}^{(1)}(p-p') + \frac{(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{q}}{m_e}\frac{\partial}{\partial p'}\delta_{\rm D}^{(1)}(p-p')\end{aligned} \tag{5.15}$$

where the second equality makes use of the fact that, for a general function $f$ of the difference of two variables, $\partial f(x-y)/\partial x = -\partial f(x-y)/\partial y$. This formal expansion is to be understood as part of the integrand over $p'$. Once we do the integral, the derivatives of delta functions will be handled by integrating by parts. With this expansion, and using the fact that $f_e(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}') \simeq f_e(\boldsymbol{q})$ (which follows from $p,\, p' \ll q$), the collision term becomes

$$C[f(\boldsymbol{p})] = \frac{\pi}{8m_e^2 p}\int\frac{d^3q}{(2\pi)^3}f_e(\boldsymbol{q})\int\frac{d^3p'}{(2\pi)^3 p'}\sum_{3\ \text{spins}}|\mathcal{M}|^2 \times\left\{\delta_{\rm D}^{(1)}(p-p') + \frac{(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{q}}{m_e}\frac{\partial\delta_{\rm D}^{(1)}(p-p')}{\partial p'}\right\}\{f(\boldsymbol{p}') - f(\boldsymbol{p})\}. \tag{5.16}$$

We now realize why we could ignore stimulated emission in the regime of interest. Including stimulated emission changes the final factor in braces to $\{f(\boldsymbol{p}')[1+f(\boldsymbol{p})] - f(\boldsymbol{p})[1+f(\boldsymbol{p}')]\}$. The additional terms in brackets simply cancel.

To proceed, we need the amplitude squared for Compton scattering. In the low-energy limit of interest to us, a textbook result[2] is that

[2] See for example exercise 11.2 in Srednicki (2007).

$$\frac{1}{2}\sum_{4\text{ spins}}|\mathcal{M}|^2 = 24\pi\sigma_{\rm T}m_e^2\left(1+[\hat{\boldsymbol{p}}\cdot\hat{\boldsymbol{p}}']^2\right) \tag{5.17}$$

where $\sigma_{\rm T}$ is the Thomson cross-section. If we are content with averaging over the polarization states of the photon with momentum $\boldsymbol{p}$ as well (hence the prefactor $1/2$), then we can use this expression for $\sum_{3\text{ spins}}|\mathcal{M}|^2$. We will in fact make one more simplification and perform an angle-average of Eq. (5.17). This turns the factor in parentheses into $4/3$, and we obtain

$$\sum_{3\text{ spins}}|\mathcal{M}|^2 = 32\pi\sigma_{\rm T}m_e^2 \qquad \text{(spin- and angle-average)}. \tag{5.18}$$

Ignoring the angular dependence changes the final collision term by a numerically subdominant contribution. It would simply distract us here, so we ignore it in the following. You can remedy this in Exercise 5.4.

By averaging over spin states of both the ingoing and outgoing photons, we are ignoring the effect of polarization of the radiation field. In reality, the amplitude for Compton scattering has a polarization dependence, which in fact leads to a small polarization of the CMB (Bond and Efstathiou, 1984; Polnarev, 1985). It turns out that the information carried by the CMB polarization spectrum is extremely valuable, which is why we will devote considerable time in Ch. 10 to understanding polarization. Compton scattering also couples polarization and temperature perturbations, so an accurate determination of the latter requires a treatment of the former. Again, we will neglect this small effect here in the derivation of the collision term.

Once we have assumed that $\sum_{\rm spins}|\mathcal{M}|^2$ is independent of the momenta involved, we can multiply out the terms in brackets in Eq. (5.16) keeping only terms first order in energy transfer. The $\boldsymbol{q}$ integral simply gives a factor of $n_e/2$ (the 2 accounting for the two spin states of the electron, i.e. $g_e = 2$), for terms that are independent of $\boldsymbol{q}$. Terms that contain a factor of $\boldsymbol{q}/m_e$, on the other hand yield $n_e\boldsymbol{u}_{\rm b}/2$ where $\boldsymbol{u}_{\rm b}$ is the bulk velocity of the electrons (the subscript "b" indicates that it is the same velocity as that of the baryons, as we will see below). So,

$$\begin{aligned}
C[f(\boldsymbol{p})] &= \frac{2\pi^2 n_e\sigma_{\rm T}}{p}\int\frac{d^3p'}{(2\pi)^3p'}\left\{\delta_{\rm D}^{(1)}(p-p')+(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{u}_{\rm b}\frac{\partial\delta_{\rm D}^{(1)}(p-p')}{\partial p'}\right\}\\
&\qquad\times\left\{f^{(0)}(p')-f^{(0)}(p)-p'\frac{\partial f^{(0)}}{\partial p'}\Theta(\hat{\boldsymbol{p}}')+p\frac{\partial f^{(0)}}{\partial p}\Theta(\hat{\boldsymbol{p}})\right\}\\
&= \frac{n_e\sigma_{\rm T}}{4\pi p}\int_0^\infty dp'p'\int d\Omega'\left[\delta_{\rm D}^{(1)}(p-p')\left(-p'\frac{\partial f^{(0)}}{\partial p'}\Theta(\hat{\boldsymbol{p}}')+p\frac{\partial f^{(0)}}{\partial p}\Theta(\hat{\boldsymbol{p}})\right)\right.\\
&\qquad\left.+(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{u}_{\rm b}\frac{\partial\delta_{\rm D}^{(1)}(p-p')}{\partial p'}(f^{(0)}(p')-f^{(0)}(p))\right],
\end{aligned} \tag{5.19}$$

where $\Omega'$ is the solid angle spanned by the unit vector $\hat{\boldsymbol{p}}'$. Here, we have only indicated the dependence of $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{p}}'$ for $\Theta$, since the dependence on $\boldsymbol{x}$, $t$ is irrelevant in the derivation of the collision term (it is always the same $\boldsymbol{x}$, $t$ since collisions are local). On the first line, we have broken up the difference $f(\boldsymbol{p}') - f(\boldsymbol{p})$ into a zeroth-order piece,[3] which cancels as expected, and a first-order part which can be neglected when multiplying the velocity term.

There are only two terms in Eq. (5.19) which depend on $\hat{\boldsymbol{p}}'$ and which therefore must be accounted for when integrating over solid angle $\Omega'$. First, there is the perturbation to the distribution function, $\Theta(\hat{\boldsymbol{p}}')$. It is convenient at this stage to introduce the *monopole*

$$\Theta_0(\boldsymbol{x}, t) \equiv \frac{1}{4\pi} \int d\Omega' \Theta(\hat{\boldsymbol{p}}', \boldsymbol{x}, t). \tag{5.20}$$

The monopole $\Theta_0$ is an integral of the photon perturbation at any given point over all photon directions. In other words, it corresponds to the fractional perturbation in the angle-averaged photon flux at a given position $\boldsymbol{x}$ and time $t$ (but phrased as a temperature perturbation via the Bose–Einstein distribution). We will later generalize this to a whole sequence of *multipole moments*, integrals of the full distribution function weighted by functions of the directions $\hat{p}$ (Eq. (5.66)). Note that, as Fig. 5.2 shows, we cannot absorb this monopole into the definition of the zeroth-order temperature since the latter is constant over all space.

The second term in Eq. (5.19) which depends on $\hat{\boldsymbol{p}}'$ is the explicit factor $\hat{\boldsymbol{p}}' \cdot \boldsymbol{u}_\mathrm{b}$. This term integrates to zero since $\boldsymbol{u}_\mathrm{b}$ is a vector that is independent of $\boldsymbol{p}$, $\boldsymbol{p}'$. Thus, the integration over solid angle leaves

$$C[f(\boldsymbol{p})] = \frac{n_e \sigma_\mathrm{T}}{p} \int_0^\infty dp' p' \left[ \delta_\mathrm{D}^{(1)}(p - p') \left( -p' \frac{\partial f^{(0)}}{\partial p'} \Theta_0 + p \frac{\partial f^{(0)}}{\partial p} \Theta(\hat{\boldsymbol{p}}) \right) + \boldsymbol{p} \cdot \boldsymbol{u}_\mathrm{b} \frac{\partial \delta_\mathrm{D}^{(1)}(p - p')}{\partial p'} (f^{(0)}(p') - f^{(0)}(p)) \right]. \tag{5.21}$$

Now the $p'$ integral can be done: in the first line by trivially integrating over the delta function and in the second by integrating by parts. We are left with

$$C[f(\boldsymbol{p})] = -p \frac{\partial f^{(0)}}{\partial p} n_e \sigma_\mathrm{T} [\Theta_0 - \Theta(\hat{\boldsymbol{p}}) + \hat{\boldsymbol{p}} \cdot \boldsymbol{u}_\mathrm{b}] \tag{5.22}$$

Already, we can anticipate the effect of Compton scattering on the photon distribution. In the absence of a bulk velocity for the electrons ($\boldsymbol{u}_\mathrm{b} = 0$), the collision term serves to drive $\Theta$ to $\Theta_0$. That is, when Compton scattering is very efficient, only the monopole perturbation survives; all anisotropies are washed out at each point in space (Fig. 5.2). Intuitively, strong scattering means that the mean free path of a photon is very small. Therefore, photons

[3] Note that we are expanding in two small quantities simultaneously: the small perturbations and the small energy transfer. Here, we are breaking up $f(\boldsymbol{p}') - f(\boldsymbol{p})$ into terms at zeroth and first order in the small perturbations.

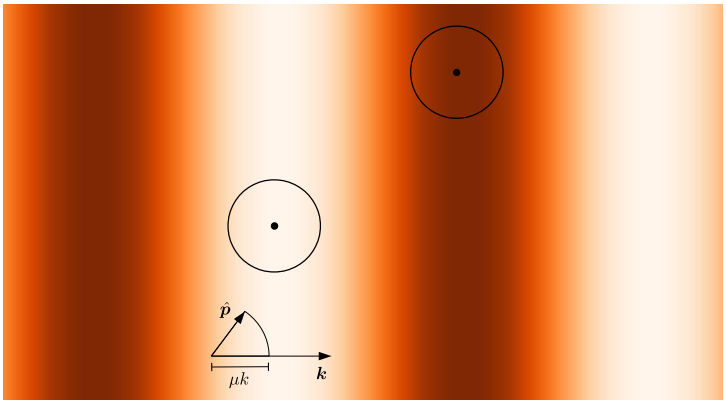



**FIGURE 5.2** A plane-wave temperature perturbation ($k$ is horizontal, as indicated) and its effect on tightly coupled photons. Dark (white) regions represent hot (cold) spots in the electron temperature. If Compton scattering is very efficient, the photons observed at a given point last scattered very nearby, within the circles which denote the last-scattering surfaces for observation points indicated by dots. The temperature on these surfaces is very close to uniform, so the distribution is almost completely described by its monopole. However, different circles (corresponding to different observers) have different temperatures due to the perturbation. So the monopole varies in space.

arriving at a given point in space last scattered off electrons that are very nearby. These nearby electrons had a temperature very similar to the local one. Therefore, photons from all directions have the same temperature, so that the flux from any direction is equal to the angle-averaged flux: $\Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, t) = \Theta_0(\boldsymbol{x}, t)$.

The situation changes slightly if the electrons carry a bulk velocity. In that case, a dipole moment will be generated in the photon distribution, determined by the amplitude and direction of the electron velocity. Still, higher-order moments such as the quadrupole vanish. Thus Compton scattering produces a photon distribution that is extremely simple to categorize: it has only a nonvanishing monopole and dipole. This is equivalent to saying that the photons behave like a fluid. Indeed, strong scattering, or *tight coupling*, produces a situation wherein the photons and electrons behave as a single fluid. Compton scattering ceases to be efficient at photon-baryon decoupling, so photons no longer behave like a fluid after recombination. However, the Boltzmann approach remains valid after decoupling and also captures the essential physics of *free streaming* that characterizes the photons' long journey from decoupling to our detectors.

## 5.3 The Boltzmann equation for photons

We can now collect the left- and right-hand sides of the Boltzmann equations from the previous two sections. A few more definitions will complete the first goal of this chapter, a linear equation for the perturbation to the photon distribution. Equating Eqs. (5.9) and

(5.22) leads to

$$\dot{\Theta} + \frac{\hat{p}^i}{a}\frac{\partial \Theta}{\partial x^i} + \dot{\Phi} + \frac{\hat{p}^i}{a}\frac{\partial \Psi}{\partial x^i} = n_e \sigma_{\rm T}[\Theta_0 - \Theta + \hat{\boldsymbol{p}} \cdot \boldsymbol{u}_{\rm b}]. \tag{5.23}$$

At this point, it is convenient to introduce the conformal time $\eta$, defined in Eq. (2.35), as our time variable. In terms of the conformal time, the Boltzmann equation becomes

$$\Theta' + \hat{p}^i \frac{\partial \Theta}{\partial x^i} + \Phi' + \hat{p}^i \frac{\partial \Psi}{\partial x^i} = n_e \sigma_{\rm T} a[\Theta_0 - \Theta + \hat{\boldsymbol{p}} \cdot \boldsymbol{u}_{\rm b}]. \tag{5.24}$$

Here, and from now on, primes will denote derivatives with respect to conformal time $\eta$, while dots continue to signify derivatives with respect to physical time $t$.

## 5.1 The virtues of Fourier space

Consider a field $\delta(\boldsymbol{x}, t)$ that obeys a linear partial differential equation, for example

$$\frac{\partial^2}{\partial t^2}\delta + f(t)\frac{\partial}{\partial t}\delta + g(t)\nabla^2\Psi = 0, \tag{5.25}$$

in terms of another field $\Psi(\boldsymbol{x}, t)$. What is noteworthy about this equation apart from its linearity is that the coefficients are functions of time $t$ only. In fact, in cosmology this property follows directly when studying small perturbations around a smooth universe: the only $\boldsymbol{x}$ dependence can be due to perturbations, and we work to linear order in them. A partial differential equation of the form Eq. (5.25) is particularly well-suited to working in Fourier space. Let us define spatial Fourier transforms through

$$\delta(\boldsymbol{x}) = \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \tilde{\delta}(\boldsymbol{k}), \tag{5.26}$$

from which follows

$$\tilde{\delta}(\boldsymbol{k}) = \int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \delta(\boldsymbol{x}). \tag{5.27}$$

Derivatives with respect to $\boldsymbol{x}$ acting on $\delta(\boldsymbol{x})$ become algebraic relations in Fourier space:

$$\frac{\partial \delta(\boldsymbol{x}, t)}{\partial x^i} \to i k_i \tilde{\delta}(\boldsymbol{k}, t). \tag{5.28}$$

Note that $k^i$ is a 3D vector in Euclidean space so that $k_i = k^i$; you do not need a factor of $g_{ij}$ to go back and forth, just as is the case for the derivatives $\partial_i$. The same goes for the velocity $u_{\rm b}^i$ and other 3-vectors. We will often characterize a mode by the magnitude of its wavevector: $k = |\boldsymbol{k}|$.

**Convention:** Throughout this book, with few exceptions, we will drop the tilde on Fourier-transformed quantities, so that, for example,

$$\tilde{\delta}(\boldsymbol{k}) \to \delta(\boldsymbol{k}). \tag{5.29}$$

This convention is used in much of the literature. Despite appearances, it is rarely confusing, because Fourier-space fields can be distinguished by their argument as in the above equation, and equations in Fourier space usually have factors of $\boldsymbol{k}$ and no spatial derivatives.

With these relations, Eq. (5.25) becomes, in Fourier space,

$$\frac{\partial^2}{\partial t^2}\delta + f(t)\frac{\partial}{\partial t}\delta - g(t)k^2\Psi = 0. \tag{5.30}$$

The partial differential equation magically turned into a set of ordinary differential equations which, moreover, are decoupled: we can solve the equation independently for each $\boldsymbol{k}$, without knowing the solution for other values of $\boldsymbol{k}'$. Another way of saying this is that every Fourier mode *evolves independently*. Instead of solving an infinite number of coupled equations, which is what Eq. (5.25) represents, we can solve for one $\boldsymbol{k}$-mode at a time. At linear order, this trick works every time in cosmology.

Eq. (5.24) is a linear partial differential equation coupling $\Theta$ to the other variables $\Phi$, $\Psi$, and $\boldsymbol{u}_\mathrm{b}$, which also evolve according to linear equations. This simplification arises because the perturbations are small. Perturbations to the CMB remain small at all cosmological epochs, while perturbations to matter are only small in the early universe. They eventually grow to become nonlinear and form collapsed objects such as galaxies, requiring new tools we will develop in Ch. 12.

As argued in Box 5.1, we should solve Eq. (5.24) in Fourier space. Before transforming to Fourier space, let us make two more definitions. First, the cosine of the angle between the wavenumber $\boldsymbol{k}$ and the photon direction $\hat{\boldsymbol{p}}$ is denoted as $\mu$:

$$\mu \equiv \frac{\boldsymbol{k}\cdot\hat{\boldsymbol{p}}}{k}. \tag{5.31}$$

From now on, $\mu$ will be the variable describing the direction of photon propagation.[4] A good way to think of $\mu$ is to go back to Fig. 5.2. The wavevector $\boldsymbol{k}$ is pointing in the direction in which the temperature is changing; it is parallel to the gradient ($\boldsymbol{k}$ is horizontal in the figure). Thus, $\Theta(k, \mu = 1)$ describes photons traveling in the direction of the gradient $\boldsymbol{k}$, along which the temperature is changing. Conversely, $\Theta(k, \mu = 0)$ describes photons traveling perpendicular to the gradient, i.e. a direction along which the temperature remains the same (vertically in the figure). While we do not distinguish Fourier-space fields in terms of notation (see Box 5.1), the appearance of $\mu$ in an equation automatically means that it is written in Fourier space.

In cosmology, velocities are generally *longitudinal*, that is, the velocities point in the same direction as $\boldsymbol{k}$:

$$\boldsymbol{u}_\mathrm{b}(\boldsymbol{k}, \eta) = \frac{\boldsymbol{k}}{k}u_\mathrm{b}(\boldsymbol{k}, \eta). \tag{5.32}$$

This is equivalent to saying that the velocity is *irrotational* (in real space, $\nabla \times \boldsymbol{u} = 0$). So, $\boldsymbol{u}_\mathrm{b}\cdot\hat{\boldsymbol{p}} = u_\mathrm{b}\mu$. Next, we define the optical depth

$$\tau(\eta) \equiv \int_\eta^{\eta_0} d\eta' n_e \sigma_\mathrm{T} a. \tag{5.33}$$

[4] We will no longer encounter the chemical potential $\mu$ here and in the following chapters.

At late times, the free electron density is small, so $\tau \ll 1$, while at early times, it is very large. Note that we have defined the limits of integration in such a way that

$$\tau' \equiv \frac{d\tau}{d\eta} = -n_e \sigma_{\rm T} a. \tag{5.34}$$

With these definitions, we are finally left with the Boltzmann equation for photons:

$$\Theta' + ik\mu\Theta + \Phi' + ik\mu\Psi = -\tau' \left[\Theta_0 - \Theta + \mu u_{\rm b}\right]. \tag{5.35}$$

Notice that different Fourier modes $\boldsymbol{k}$ are decoupled: we can solve for each value of $k$ and $\mu$ independently.

## 5.4 The Boltzmann equation for cold dark matter

The derivation of the Boltzmann equation for any other constituent in the universe proceeds by repeating the same steps as we did for the photons. Of particular importance is the evolution of the dark matter. In all viable models of structure formation, dark matter plays an important role in the growth of structure through its gravitational effect. As in the case of photons, the Boltzmann equation is the correct starting point for describing the evolution of dark matter.

There are several ways in which the dark matter distribution differs from that of the photons. First, at epochs long after its production, dark matter by definition does not interact with any of the other constituents in the universe. Thus we need not deal with any collision terms.[5] Second, cold dark matter, in contrast to the photons, is nonrelativistic; typical velocities of dark matter particles are much less than the speed of light.

Thus, the appropriate version of the Boltzmann equation to use is the collisionless Boltzmann equation for massive particles, Eq. (3.76) that we derived in Sect. 3.3.4:

$$\frac{\partial f_{\rm c}}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f_{\rm c}}{\partial x^i} - \left[H + \dot{\Phi} + \frac{E}{ap}\hat{p}^i \Psi_{,i}\right] p \frac{\partial f_{\rm c}}{\partial p} = 0. \tag{5.36}$$

The main difference from the corresponding equation for radiation is the presence of factors of $p/E$, or velocity. For dark matter particles, these velocity factors suppress any free streaming, as we will shortly see.

In the massless case, to proceed further we used our knowledge of the distribution function. Namely, we knew that the zeroth-order distribution function was Bose–Einstein, and we perturbed around this zeroth-order solution. For cold dark matter particles, we do not need such detailed information about the zeroth-order distribution function. All we need to know is that these particles are very nonrelativistic. Therefore, increasingly higher-order powers of $p$ will be negligible: only terms with $(p/m)^0$ and $(p/m)^1$ need be retained. That

[5] Unless the dark matter is strongly self-interacting, a scenario which has gained attention in recent years (Tulin and Yu, 2018). We will not consider this possibility here though.

is, we will neglect terms of order $(p/m)^2$. This means that we include the bulk motion of dark matter, but not its velocity dispersion, and hence treat the dark matter as an effective fluid. We return to a more detailed justification, and limits, of this approximation in Ch. 12.

Then, instead of assuming a form for $f_{\rm c}$, we take moments of Eq. (3.76) (see Exercise 3.8). First, multiply both sides by the phase space volume $d^3p/(2\pi)^3$ and integrate. This leads to

$$\frac{\partial}{\partial t}\int\frac{d^3p}{(2\pi)^3}f_{\rm c}+\frac{1}{a}\frac{\partial}{\partial x^i}\int\frac{d^3p}{(2\pi)^3}f_{\rm c}\frac{p\hat{p}^i}{E(p)}-\left[H+\dot{\Phi}\right]\int\frac{d^3p}{(2\pi)^3}p\frac{\partial f_{\rm c}}{\partial p}$$
$$-\frac{1}{a}\frac{\partial\Psi}{\partial x^i}\int\frac{d^3p}{(2\pi)^3}\frac{\partial f_{\rm c}}{\partial p}E(p)\hat{p}^i=0. \tag{5.37}$$

Note that, since they are independent variables, the integral over $p$ passes through the partial derivatives with respect to $x^i$ and $t$. Integration by parts shows that the last term vanishes. The remainder of the terms are all relevant, though. To simplify, let us recall that the dark matter density is[6]

$$n_{\rm c}=\int\frac{d^3p}{(2\pi)^3}f_{\rm c}, \tag{5.38}$$

while the fluid velocity is defined as

$$u^i_{\rm c}\equiv\frac{1}{n_{\rm c}}\int\frac{d^3p}{(2\pi)^3}f_{\rm c}\frac{p\hat{p}^i}{E(p)}. \tag{5.39}$$

Notice that we use the notation $\boldsymbol{u}$ for *fluid* velocities, to be distinguished from the velocities of individual *particles*. It is important to remember the physical distinction between the two: the former describes the bulk motion of matter averaged over many particles, and could be much smaller than the individual particle velocities (although not in the case of cold dark matter).

The first two terms in Eq. (5.37), then, can be simply expressed in terms of the velocity and the density. The third term can be integrated by parts:

$$\begin{aligned}\int\frac{d^3p}{(2\pi)^3}p\frac{\partial f_{\rm c}}{\partial p}&=\frac{1}{(2\pi)^3}\int_0^\infty dp\,p^3\frac{\partial}{\partial p}\int d\Omega\,f_{\rm c}\\&=-3\frac{1}{(2\pi)^3}\int_0^\infty dpp^2\int d\Omega\,f_{\rm c}\\&=-3n_{\rm c}.\end{aligned} \tag{5.40}$$

So the zeroth moment of the Boltzmann equation leads to the cosmological generalization of the continuity equation:

$$\frac{\partial n_{\rm c}}{\partial t}+\frac{1}{a}\frac{\partial(n_{\rm c}u^i_{\rm c})}{\partial x^i}+3\left[H+\dot{\Phi}\right]n_{\rm c}=0. \tag{5.41}$$

[6] Here we have incorporated the spin degeneracy $g_{\rm c}$ into the phase space distribution $f_{\rm c}$ to avoid irrelevant factors of $g_{\rm c}$ throughout the derivation.

The first two terms here are the standard continuity equation from fluid mechanics. The last term arises due to the FLRW metric and its perturbations, in particular the dilution by the expansion of space (recall that $H + \dot{\Phi}$ corresponds to the local perturbed Hubble rate).

To go further, we can separate zeroth-order and first-order terms in Eq. (5.41). The velocity is of first order as is $\Phi$, so the only zeroth-order terms are

$$\frac{\partial \bar{n}_{\rm c}}{\partial t} + 3H\bar{n}_{\rm c} = 0 \tag{5.42}$$

where $\bar{n}_{\rm c}$ is the zeroth-order, homogeneous part of the density. Equivalently, we have

$$\frac{d(\bar{n}_{\rm c}a^3)}{dt} = 0 \Rightarrow \bar{n}_{\rm c} \propto a^{-3}, \tag{5.43}$$

a relation we proved in Ch. 2 by using the conservation of the energy-momentum tensor.

Now let us extract the first-order part of Eq. (5.41). All factors of $n_{\rm c}$ multiplying the first-order quantities $u_c$ and $\Phi$ may be set to $\bar{n}_{\rm c}$. Everywhere else, we need to expand $n_{\rm c}$ out to include a first-order perturbation. In particular, we will set

$$n_{\rm c}(\boldsymbol{x}, t) = \bar{n}_{\rm c}(t)[1 + \delta_{\rm c}(\boldsymbol{x}, t)] \tag{5.44}$$

which defines the first-order piece as $\bar{n}_{\rm c}\delta_{\rm c}$. Since the energy density of matter is equal to mass times $n_{\rm c}$, $\delta_{\rm c}$ is also the fractional overdensity, $\delta\rho_{\rm c}/\rho_{\rm c}$, of the dark matter. After dividing by $\bar{n}_{\rm c}$, the first-order equation is therefore

$$\frac{\partial \delta_{\rm c}}{\partial t} + \frac{1}{a}\frac{\partial u^i_{\rm c}}{\partial x^i} + 3\dot{\Phi} = 0. \tag{5.45}$$

As it stands, we have introduced two new perturbation variables for the dark matter, the density perturbation $\delta_{\rm c}$ and the velocity $\boldsymbol{u}_{\rm c}$. Eq. (5.45) is only one equation, though, for these two variables. We need another. To get it, we return to the unintegrated Boltzmann equation (5.36). We have just taken its zeroth moment; to extract a second equation, let us take its first moment. In particular, multiply Eq. (5.36) by $(d^3p/(2\pi)^3)\, p\hat{p}^j/E$ and then integrate. The first moment equation is then

$$\begin{aligned}\frac{\partial}{\partial t}\int \frac{d^3p}{(2\pi)^3} f_{\rm c}\frac{p\hat{p}^j}{E} + \frac{1}{a}\frac{\partial}{\partial x^i}\int \frac{d^3p}{(2\pi)^3} f_{\rm c}\frac{p^2}{E^2(p)}\hat{p}^i\hat{p}^j - \left[H + \dot{\Phi}\right]\int \frac{d^3p}{(2\pi)^3}\frac{\partial f_{\rm c}}{\partial p}\frac{p^2\hat{p}^j}{E} \\ -\frac{1}{a}\frac{\partial \Psi}{\partial x^i}\int \frac{d^3p}{(2\pi)^3}\frac{\partial f_{\rm c}}{\partial p} p\hat{p}^i\hat{p}^j = 0.\end{aligned} \tag{5.46}$$

The first two terms are straightforward: the first is the time derivative of $n_{\rm c}u^i_{\rm c}$ while the second can be safely neglected since it is of order $(p/E)^2$. The last two integrals must be handled more carefully, though, because of the derivatives acting on $f_{\rm c}$. Let us do the inte-

gration by parts explicitly in the third term. The integral is

$$\int \frac{d^3p}{(2\pi)^3}\frac{\partial f_{\rm c}}{\partial p}\frac{p^2\hat{p}^j}{E} = \int \frac{d\Omega}{(2\pi)^3}\hat{p}^j\int_0^\infty dp\,\frac{p^4}{E}\frac{\partial f_{\rm c}}{\partial p}$$
$$= -\int \frac{d\Omega}{(2\pi)^3}\hat{p}^j\int_0^\infty dp f_{\rm c}\left(\frac{4p^3}{E}-\frac{p^5}{E^3}\right). \tag{5.47}$$

The first term, $\propto -4p^3/E$, yields $-4n_{\rm c}u_{\rm c}^j$ upon integration, while the term involving $p^5/E^3 = (p^2/E^2)(p^3/E)$ is negligible following our counting. The same steps carry through for the last term in Eq. (5.46); the one additional fact we need is that

$$\int d\Omega\hat{p}^i\hat{p}^j = \delta^{ij}\frac{4\pi}{3}. \tag{5.48}$$

So the first moment of the Boltzmann equation is

$$\frac{\partial(n_{\rm c}u_{\rm c}^j)}{\partial t} + 4Hn_{\rm c}u_{\rm c}^j + \frac{n_{\rm c}}{a}\frac{\partial\Psi}{\partial x^j} = 0. \tag{5.49}$$

This equation has no zeroth-order parts, since the velocity is a first-order quantity. Therefore, we need to extract only the first-order terms, which allows us to set $n_{\rm c}\to\bar{n}_{\rm c}$ everywhere. Using the time dependence we found in Eq. (5.43) we arrive at

$$\frac{\partial u_{\rm c}^j}{\partial t} + Hu_{\rm c}^j + \frac{1}{a}\frac{\partial\Psi}{\partial x^j} = 0. \tag{5.50}$$

Eq. (5.45) and Eq. (5.50) are the two equations governing the evolution of the density and the velocity of the cold dark matter. The momentum conservation, or Euler equation (5.50) does not have the standard $(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}$ term, since any term with two factors of $\boldsymbol{u}$ is manifestly of second order (this term will appear in Ch. 12 when we study dark matter beyond linear perturbations). An interesting feature of the two equations is generic to this process of integrating the Boltzmann equations to get the fluid equations: the integrated Boltzmann equation for the $l$th moment depends on the moment of order $l+1$; e.g. the equation for the density (zeroth moment of the distribution function) depends on the velocity (first moment). This process of integrating leads to an infinite hierarchy of equations for the moments of the distribution function. Indeed, we will see that this is one way of solving the Boltzmann equation for the photons, Eq. (5.35). In the case of CDM, we have closed the hierarchy by setting the next, second, moment to zero, following our assumption that the dark matter is *cold*. Specifically, we have dropped all terms of order $(p/E)^2$ and higher. Thus, Eq. (5.45) and Eq. (5.50) are a closed set of equations for the cold dark matter distribution.[7] For particles with larger velocities, such as massive neutrinos, the hierarchy cannot be simply closed in this way, and we need to keep higher moments.

[7] We still need equations for the gravitational potentials $\Phi$ and $\Psi$. These come from Einstein's equations, which are the topic of the next chapter.

Let us finally rewrite Eq. (5.45) and Eq. (5.50) in terms of conformal time $\eta$ and in Fourier space. The continuity equation becomes

$$\delta_{\rm c}{}' + iku_{\rm c} + 3\Phi' = 0. \tag{5.51}$$

where again we have assumed that the velocity is irrotational so $u_{\rm c}^i = (k^i/k)u_{\rm c}$. The Euler equation is

$$u_{\rm c}{}' + \frac{a'}{a}u_{\rm c} + ik\Psi = 0. \tag{5.52}$$

This equation at least partly justifies our assumption that $\boldsymbol{u}_{\rm c}$ is irrotational, since it says that $\boldsymbol{u}_{\rm c}$ is sourced by the gradient of a scalar potential $\Psi$. Any curl component would have to be set in the initial conditions for the velocity.

## 5.5 The Boltzmann equation for baryons

The next component of the universe that requires a set of Boltzmann equations are the electrons and protons. These components are often grouped together and called *baryons*. This obvious misnomer is motivated by the fact that the energy density of these coupled particles is dominated by the rest masses of the protons and neutrons making up the hydrogen and helium nuclei. In the following, we will simply speak of protons although one should keep in mind that this includes helium as well as trace amounts of heavier nuclei.

Electrons and protons are coupled by Coulomb scattering ($e + p \leftrightarrow e + p$). The Coulomb scattering rate is much larger than the expansion rate at all epochs of interest. This tight coupling forces the electron and proton overdensities to a common value:

$$\frac{\rho_e - \bar{\rho}_e}{\bar{\rho}_e} = \frac{\rho_p - \bar{\rho}_p}{\bar{\rho}_p} \equiv \delta_{\rm b} \tag{5.53}$$

where we bow to common usage with the subscript b. Similarly, the velocities of the two species are forced to a common value,

$$\boldsymbol{u}_e = \boldsymbol{u}_p \equiv \boldsymbol{u}_{\rm b}. \tag{5.54}$$

After recombination, when electrons and nuclei first form atoms, this tight coupling remains, while the neutral atoms are now decoupled from the photons. Because of their tight mutual coupling and correspondingly small mean free path, and the fact that they are nonrelativistic since $T \ll m_e$, electrons and nuclei can be treated as a nonrelativistic fluid and we only have to take the first two moments of their Boltzmann equation, as we did for dark matter. However, we will need to keep track of the coupling to photons via Compton scattering.

The procedure of taking moments of the Boltzmann equation to derive equations for $\delta_{\rm b}$ and $u_{\rm b}$ then proceeds just as in the case of CDM, for the left-hand-side at least. The zeroth

moment that led us to Eq. (5.51) for CDM correspondingly yields

$$\delta_{\rm b}' + iku_{\rm b} + 3\Phi' = 0, \tag{5.55}$$

after switching to Fourier space and conformal time. Here we have put the right-hand side to zero, despite collisions, since the collisions preserve the number of electrons and protons. This holds for Coulomb scattering, where $e^- + N \leftrightarrow e^- + N$, as well as for Compton scattering $e^- + \gamma \leftrightarrow e^- + \gamma$. At the epochs we will be interested in, around and after recombination, reactions that change the number of electrons and nucleons such as pair production, annihilation and nuclear reactions are irrelevant. The continuity equation with vanishing source term precisely captures this number conservation.

The second equation for the baryons is obtained by taking the first moments of the Boltzmann equations for electrons and baryons and adding them together. We did something similar for the dark matter; there we first multiplied by $\boldsymbol{p}/E$ and then integrated over all momenta. Here we do the same but without the $1/E$ factor. Since all particles involved are nonrelativistic, our results from the dark matter case carry over as long as we multiply them by a factor of $m$. The left-hand side of the integrated electron equation, for example, will look exactly like the left-hand side of Eq. (5.49) except it will be multiplied by $m_e$. The proton equation will be multiplied by $m_p$. Since the proton mass is so much larger than the electron mass, the sum of the two left-hand sides will be dominated by the protons. So, following Eq. (5.49), we have

$$m_p \frac{\partial(n_{\rm b} u_{\rm b}^j)}{\partial t} + 4Hm_p n_{\rm b} u_{\rm b}^j + \frac{m_{\rm p} n_{\rm b}}{a}\frac{\partial \Psi}{\partial x^j} = F_{e\gamma}^j(\boldsymbol{x}, t). \tag{5.56}$$

This time, the collision term $\boldsymbol{F}_{e\gamma}$ does not vanish. The first moment of the Boltzmann equation describes momentum conservation. While the number of electrons and nuclei is preserved, their momentum is not, since Compton scattering transfers momentum between photons and electrons, captured by the force density $\boldsymbol{F}_{e\gamma}$. The electrons in turn transfer it immediately to the nuclei.[8] Dividing both sides by $\rho_{\rm b} = m_p \bar{n}_{\rm b}$, we are left with

$$\frac{\partial u_{\rm b}^j}{\partial t} + Hu_{\rm b}^j + \frac{1}{a}\frac{\partial \Psi}{\partial x^j} = \frac{1}{\rho_{\rm b}} F_{e\gamma}^j(\boldsymbol{x}, t). \tag{5.57}$$

Again, so far we have followed the same steps as in the derivation of the first moment of the dark matter Boltzmann equation.

The final step is to evaluate the integrated collision term on the right-hand side, for which we will use a convenient shortcut. We have argued that $\boldsymbol{F}_{e\gamma}$ describes the momentum transfer between photons and electrons. Since momentum is conserved in each scattering event, this force term has to be precisely equal and opposite to the force term appearing in the photon analog of the baryon Euler equation. To get this equation, we

[8] In principle, photons also scatter off the nuclei, but this interaction is suppressed so strongly, by $m_e^2/m_p^2 < 10^{-6}$, that it is entirely sufficient to include only the electron scattering term.

have to take the first moment of the photon Boltzmann equation, specifically the photon collision term Eq. (5.22).

First, let us switch to Fourier space. Anticipating that the direction of the force term $\boldsymbol{F}_{e\gamma}$ will be aligned with the wavevector $\boldsymbol{k}$, we multiply the Fourier-space version of Eq. (5.22) by $\hat{k}^j$ before taking the first moment. In addition, since the momentum $n_e u_e^i$ of electrons counts both spin states, i.e. it is twice the first moment of the distribution function, we have to multiply the collision term by a factor of 2. Together with the minus sign from momentum conservation, this implies that we multiply Eq. (5.22) by $-2p\mu$ and integrate over $\boldsymbol{p}$, the photon momentum:

$$\begin{aligned}\frac{1}{\rho_\mathrm{b}}\hat{k}_i F^i_{e\gamma}(\boldsymbol{x},t) &= -\frac{2n_e\sigma_\mathrm{T}}{\rho_\mathrm{b}}\int\frac{d^3p}{(2\pi)^3}p\mu\left[-p\frac{\partial f^{(0)}}{\partial p}\right]\left[\Theta_0-\Theta(\mu)+\mu u_\mathrm{b}\right]\\ &= \frac{2n_e\sigma_\mathrm{T}}{\rho_\mathrm{b}}\int_0^\infty\frac{dp}{2\pi^2}p^4\frac{\partial f^{(0)}}{\partial p}\int_{-1}^{1}\frac{d\mu}{2}\mu\left[\Theta_0-\Theta(\mu)+\mu u_\mathrm{b}\right]. \end{aligned} \tag{5.58}$$

The integral over $p$ can be done by integrating by parts: it is $-2\rho_\gamma$, since the background energy density of photons $\rho_\gamma$ is twice the momentum integral over $pf^{(0)}(p)$. The $\mu$-integration over the first and third terms is straightforward—the first term vanishes, while the third gives $u_\mathrm{b}/3$. The integral over the second term yields the first moment of the perturbation $\Theta$. Recall the definition of the zeroth moment $\Theta_0$ of the photon distribution in Eq. (5.20). It makes sense therefore to define the first moment, the *dipole*, as

$$\Theta_1(k,\eta)\equiv i\int_{-1}^{1}\frac{d\mu}{2}\mu\Theta(\mu,k,\eta) \tag{5.59}$$

where the factor of $i$ follows convention. So Eq. (5.58) becomes

$$\frac{1}{\rho_\mathrm{b}}\hat{k}_i F^i_{e\gamma}(\boldsymbol{x},t) = -n_e\sigma_\mathrm{T}\frac{4\rho_\gamma}{\rho_\mathrm{b}}\left[i\Theta_1+\frac{1}{3}u_\mathrm{b}\right]. \tag{5.60}$$

In order to see why the dipole of the radiation field appears in the Euler equation for the baryons, recall that what matters is a net momentum transfer between the photons and electrons. In an isotropic radiation field, no net momentum transfer will happen. On the other hand, if there is a dipole, then more energetic photons come from one direction than from the opposite direction. Electrons moving in the direction of the higher temperature will be facing a headwind, resulting in a drag force pointing in the opposite direction. This effect is known as *Compton drag*. More precisely, $F^i_{e\gamma}$ is a force density (analogous to a pressure gradient) exerted by scattering of photons off the electrons. The force density is given by the collision rate ($n_e\sigma_\mathrm{T}$) multiplied by the mean momentum transfer in each collision, which is of order of the photon density times the dipole of the photon distribution.

We now have an expression for the collision term which can be inserted into Eq. (5.57), and, after switching to conformal time, we have

$$u_\mathrm{b}{}' + \frac{a'}{a}u_\mathrm{b} + ik\Psi = \tau'\frac{4\rho_\gamma}{3\rho_\mathrm{b}}\left[3i\Theta_1+u_\mathrm{b}\right]. \tag{5.61}$$

Why is there a factor of the baryon energy density in the denominator? That is, since photons scatter primarily off electrons, why does the total baryon density (which is dominated by protons) appear in this velocity equation? Physically, it arises from the fact that moving electrons is difficult because they are tightly coupled to protons via Coulomb scattering. If the proton was infinitely heavy, so $\rho_\mathrm{b} \to \infty$, Compton scattering would not change the electron velocity at all; it would not have any impact on the combined proton-electron fluid. We derived Eq. (5.61) by setting $n_e = n_p = n_\mathrm{b}$, that is, we assumed that hydrogen is completely ionized. However, the result turns out to be valid even if there is an appreciable amount of neutral hydrogen, so that $n_e < n_\mathrm{b}$. Indeed after recombination, most protons are bound in neutral hydrogen atoms. And even before recombination, a small fraction are in helium atoms or ions. However, even neutral hydrogen and helium are tightly coupled to electrons and protons (see Exercise 5.6), so our result Eq. (5.61) indeed describes all baryons.

## 5.6 The Boltzmann equation for neutrinos

Finally, we turn to the remaining abundant species of particles, the neutrinos, with distribution function $f_\nu(\boldsymbol{x}, \boldsymbol{p}, t)$. Let us proceed in analogy to the photons, since the neutrinos follow an equilibrium distribution with a temperature $T_\nu(a)$ at zeroth order (see Sect. 2.4.4 and Exercise 3.9), and they are relativistic in the early universe. So we can again phrase the perturbation to their distribution function in terms of a temperature perturbation, denoted $\mathcal{N}(\boldsymbol{x}, \boldsymbol{p}, \eta)$, just as we did for the photons in Sect. 5.1; that is, we write

$$\begin{aligned} f_\nu(\boldsymbol{x}, \boldsymbol{p}, t) &= \left[\exp\left\{\frac{p}{T_\nu(t)[1+\mathcal{N}(\boldsymbol{x}, \hat{\boldsymbol{p}}, t)]}\right\} + 1\right]^{-1} \\ &= \left[1 - \mathcal{N}(\boldsymbol{x}, \boldsymbol{p}, t) p \frac{d}{dp}\right] f_\nu^{(0)}(p), \end{aligned} \tag{5.62}$$

where $f_\nu^{(0)}(p) = [e^{p/T_\nu(a)} + 1]^{-1}$ is the zeroth-order neutrino distribution and the second line expands to linear order in $\mathcal{N}$. During the epochs of interest, that is, from neutrino decoupling onward, any non-gravitational interactions of neutrinos are negligible, so the appropriate Boltzmann equation is the collisionless Boltzmann equation for massive particles, Eq. (3.76):

$$\frac{df_\nu}{dt} = \frac{\partial f_\nu}{\partial t} + \frac{p}{E_\nu(p)} \frac{\hat{p}^i}{a} \frac{\partial f_\nu}{\partial x^i} - \left[H + \dot{\Phi} + \frac{E_\nu(p)}{ap} \hat{p}^i \Psi_{,i}\right] p \frac{\partial f_\nu}{\partial p} = 0. \tag{5.63}$$

Inserting Eq. (5.62), the zeroth-order terms cancel (by construction $f_\nu^{(0)}$ obeys the homogeneous Boltzmann equation), and we obtain at first order

$$\frac{\partial \mathcal{N}}{\partial t} + \frac{p}{E_\nu(p)} \frac{\hat{p}^i}{a} \frac{\partial \mathcal{N}}{\partial x^i} - Hp \frac{\partial \mathcal{N}}{\partial p} + \dot{\Phi} + \frac{E_\nu(p)}{ap} \hat{p}^i \Psi_{,i} = 0. \tag{5.64}$$

Now we convert to conformal time derivatives and move to Fourier space, yielding

$$\mathcal{N}'(\boldsymbol{k}, p, \mu, \eta) + ik\mu\frac{p}{E_\nu(p)}\mathcal{N} - Hp\frac{\partial}{\partial p}\mathcal{N} = -\Phi' - ik\mu\frac{E_\nu(p)}{p}\Psi, \tag{5.65}$$

which is our desired first-order Boltzmann equation for neutrinos. Apart from the absent collision term, it differs from that for the photons through the factors of $p/E_\nu(p)$, which reduce to unity when the neutrinos are relativistic. At late times, when the temperature drops below $m_\nu$, the first factor of $p/E$ on the left reflects the slow-down in free streaming due to the sluggish massive neutrinos. The factor of $E/p$ in the final term reflects the fact that slow-moving neutrinos spend more time in potential wells and hence their motion will be more influenced by them.

An important point to notice is that we can no longer assume that $\mathcal{N}$ is independent of $p$, i.e. that it only depends on $\hat{\boldsymbol{p}}$, $\boldsymbol{x}$, $\eta$; hence the additional third term in Eq. (5.65) which does not appear in the equation for $\Theta$. This is because neutrinos in different parts of the distribution move differently once they are no longer ultra-relativistic. For example, $p/E_\nu(p)$ can be very different for neutrinos in the low-energy tail of the distribution than for those in the high-energy tail. For the same reason, we need different distribution functions for the different neutrino mass states. Fortunately, if we are only interested in the behavior of neutrinos up to recombination, then we can set $p/E_\nu(p) = 1$ in Eq. (5.65), and neglect the $p$-dependence of $\mathcal{N}$, reducing it to the collisionless version of the photon Boltzmann equation; then, we can further describe all three neutrino generations with a single $\mathcal{N}$. We will make use of this simplification in later chapters, but will also learn that including neutrino masses becomes very important in the late universe.

## 5.7 Summary

The constituents of the universe are not distributed completely uniformly in space—the only exception to this is the cosmological constant, which we have consequently ignored in this chapter. To describe the evolution of the perturbations, we have to solve the perturbed Boltzmann equations which we derived in this chapter. For the nonrelativistic components such as the dark matter and the baryons, the Boltzmann equation can be simplified significantly by taking moments in terms of the particle momentum, and keeping only the lowest-order moments: the overdensity $\delta_{\rm c}(\boldsymbol{x}, t)$ ($\delta_{\rm b}(\boldsymbol{x}, t)$) and the velocity $\boldsymbol{u}_{\rm c}(\boldsymbol{x}, t)$ ($\boldsymbol{u}_{\rm b}(\boldsymbol{x}, t)$) for dark matter (baryons). As explained in Box 5.1, it is convenient to transform the linear evolution equations to Fourier space. Then, the evolution of a mode associated with wavevector $\boldsymbol{k}$ is independent of any other Fourier modes. Further, we will find it convenient to use conformal time $\eta$ as the evolution variable. To summarize, we have equations for $\delta_{\rm c}(\boldsymbol{k}, \eta)$, $\delta_{\rm b}(\boldsymbol{k}, \eta)$, $u_{\rm c}(\boldsymbol{k}, \eta)$, and $u_{\rm b}(k, \eta)$. The scalar velocities here are the components parallel to $\boldsymbol{k}$; these are the only ones that are cosmologically relevant.

Relativistic particles such as photons and neutrinos require more information to characterize. They have not only a monopole perturbation (the equivalent of $\delta_{\rm c}$) and a dipole (the equivalent of a velocity), but also a quadrupole, octopole, and higher moments as

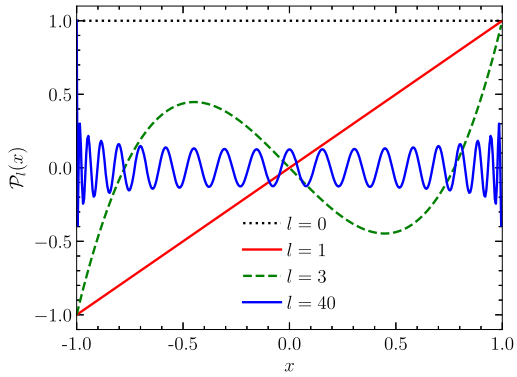



**FIGURE 5.3** Some Legendre polynomials. Note that the higher-order ones vary on smaller scales than the low-order ones do. In general $\mathcal{P}_l$ crosses zero $l$ times between $-1$ and 1.

well. In other words, the photon distribution depends not only on $\boldsymbol{x}$ and time but also on the direction of propagation of the photon, $\hat{\boldsymbol{p}}$. In Fourier space, therefore, the photon perturbations depend not only on $k$ and $\eta$ but also on $\hat{\boldsymbol{p}} \cdot \hat{\boldsymbol{k}}$, which we defined as $\mu$. Thus, the photon perturbation variable is $\Theta(k, \mu, \eta)$. Neutrino perturbations require a separate variable, $\mathcal{N}(k, p, \mu, \eta)$ (in fact, one for each mass eigenstate), with an additional dependence on $p$ due to the fact that neutrinos have mass. Fortunately, we can neglect neutrino masses and the $p$-dependence of $\mathcal{N}$ at least up until recombination.

We found it useful to define the monopole (Eq. (5.20)) and dipole (Eq. (5.59)) of the photon distribution. We know that these moments, $\Theta_0(k, \eta)$ and $\Theta_1(k, \eta)$, do not completely characterize the photon distribution. Instead, we need the general, $l$th multipole moment of the temperature field, defined as

$$\Theta_l(k, \eta) \equiv \frac{1}{(-i)^l} \int_{-1}^{1} \frac{d\mu}{2} \mathcal{P}_l(\mu) \Theta(\mu, k, \eta), \tag{5.66}$$

where $\mathcal{P}_l$ is the Legendre polynomial of order $l$ (Eq. (C.2)). The quadrupole corresponds to $l = 2$, octopole to $l = 3$, etc. The higher Legendre polynomials have structure on smaller scales (see Fig. 5.3), so the higher moments capture information about the small-scale anisotropies of the radiation field. The photon perturbations can be described either by $\Theta(k, \mu, \eta)$ or by a hierarchy of moments, $\Theta_l(k, \eta)$. The same multipole expansion can be applied to the neutrino distribution.

We have postponed a discussion of polarization until Ch. 10, but mentioned in Sect. 5.2 that a completely accurate treatment of anisotropies in the temperature requires an incorporation of polarization effects. Again, waiting until Ch. 10 for more formal definitions, let us call the photon polarization field $\Theta_P$. Upon Fourier transforming, it too depends on $k$, $\mu$, and $\eta$, and we denote its Legendre multipole decomposition as $\Theta_{P,l}$.

We now collect the equations we have derived for the photons, dark matter, baryons, and neutrinos:

$$\Theta' + ik\mu\Theta = -\Phi' - ik\mu\Psi - \tau'\left[\Theta_0 - \Theta + \mu u_{\rm b} - \frac{1}{2}\mathcal{P}_2(\mu)\Pi\right] \tag{5.67}$$

with

$$\Pi = \Theta_2 + \Theta_{P,2} + \Theta_{P,0}. \tag{5.68}$$

Eq. (5.67) is the Boltzmann equation for photons we have derived, with one change, the last term $\mathcal{P}_2\Pi/2$, which requires some explanation. First, note that it is proportional to the second Legendre polynomial, $\mathcal{P}_2(\mu) = (3\mu^2 - 1)/2$. From Eq. (5.68), one of the new terms then is $\mathcal{P}_2\Theta_2/2$; this term accounts for the angular dependence of Compton scattering, which we ignored in Sect. 5.2. The other parts of $\Pi$ represent the fact that the temperature field is also coupled to the polarization field $\Theta_P$. While we postpone the discussion of the equation for $\Theta_P$, it is worth mentioning that $\Theta_P$ is sourced only by the quadrupole $\Theta_2$ of the temperature distribution, and none of the other temperature moments.

The remaining equations are

$$\delta_{\rm c}{}' + iku_{\rm c} = -3\Phi', \tag{5.69}$$

$$u_{\rm c}{}' + \frac{a'}{a}u_{\rm c} = -ik\Psi, \tag{5.70}$$

$$\delta_{\rm b}' + iku_{\rm b} = -3\Phi', \tag{5.71}$$

$$u_{\rm b}{}' + \frac{a'}{a}u_{\rm b} = -ik\Psi + \frac{\tau'}{R}[u_{\rm b} + 3i\Theta_1], \tag{5.72}$$

$$\mathcal{N}' + ik\mu\frac{p}{E_\nu(p)}\mathcal{N} - Hp\frac{\partial}{\partial p}\mathcal{N} = -\Phi' - ik\mu\frac{E_\nu(p)}{p}\Psi. \tag{5.73}$$

In the equation for the baryon velocity (5.72), the ratio of photon to baryon density has been replaced by $R$, defined as[9]

$$\frac{1}{R(\eta)} \equiv \frac{4\rho_\gamma(\eta)}{3\rho_{\rm b}(\eta)}. \tag{5.74}$$

The derivations in this chapter are based on the seminal paper of Ma and Bertschinger (1995). While it skips many of the steps presented here, we highly recommend it as further reading (it also has the added virtue of equations in both conformal-Newtonian and synchronous gauges; see the exercise below).

[9] Not to be confused with the ratio of photon to baryon *number*, $\eta_{\rm b}$, which is constant with time and much smaller!

## Exercises

**5.1** The metric in synchronous gauge is

$$\begin{aligned} g_{00}(\boldsymbol{x},t) &= -1, \\ g_{0i}(\boldsymbol{x},t) &= 0, \\ g_{ij}(\boldsymbol{x},t) &= a^2(t)[\delta_{ij} + h_{ij}(\boldsymbol{x},t)], \end{aligned} \tag{5.75}$$

with perturbations in Fourier space given by

$$\tilde{h}_{ij}(\boldsymbol{k} = k\hat{\boldsymbol{e}}_z, t) = \begin{pmatrix} -2\tilde{\mathfrak{n}}(\boldsymbol{k},t) & 0 & 0 \\ 0 & -2\tilde{\mathfrak{n}}(\boldsymbol{k},t) & 0 \\ 0 & 0 & \tilde{h}(\boldsymbol{k},t) + 4\tilde{\mathfrak{n}}(\boldsymbol{k},t) \end{pmatrix}. \tag{5.76}$$

In this exercise, we keep the Fourier-space tilde explicit. Here we have chosen the wavevector $\boldsymbol{k}$ to lie in the $z$ direction. Following the steps in Sect. 3.3, derive the equivalent of Eq. (5.35) in synchronous gauge:

$$\tilde{\Theta}' + ik\mu\tilde{\Theta} + \frac{1}{2}\mu^2\tilde{h}' + 2\mathcal{P}_2(\mu)\tilde{\mathfrak{n}}' = -\tau'\left[\tilde{\Theta}_0 - \tilde{\Theta} + \mu u_{\rm b}\right]. \tag{5.77}$$

**5.2** Start from the zeroth-order unintegrated Boltzmann Eq. (5.5). Integrate this equation over all momenta to show that the number density falls off as $a^3$.

**5.3** Show that the Pauli blocking factor $1 - f_e$ can be set to 1 at all times from BBN through recombination. First find $f_e$, which depends on $T_e$ and $\mu_e$, as a function of temperature and number density using the results and approximations of Sect. 4.1 (i.e. assume that $T_e \ll m_e$). Then, show that in this regime $f_e$ is much less than 1.

**5.4** Account for the angular dependence of Compton scattering. Start from Eq. (5.16) but instead of using the angle-averaged amplitude-squared Eq. (5.18), take the correct polarization-averaged expression Eq. (5.17). Show that accounting for the angular dependence introduces the factor of $(1/2)\mathcal{P}_2(\mu)\Theta_2$ presented in Eq. (5.67).

**5.5** Derive the continuity equation by using the Boltzmann equation, rederiving the results of Sect. 2.3 that were based on the energy-momentum tensor. Multiply the zeroth-order part of Eq. (3.76) by $d^3p/(2\pi)^3\, E(p)$ and integrate. Show that the resulting equation is identical to Eq. (2.56).

**5.6** Show that electrons, nuclei, and atoms are tightly coupled all the way through recombination.

**(a)** Compute the ratio of the Coulomb scattering rate to the Hubble rate. You may assume that all electrons and protons are ionized.

**(b)** Show that the rate for neutral hydrogen to scatter off free protons is always much larger than the expansion rate even when the ionization fraction is on the order of $10^{-4}$ (cf. Fig. 4.4).

# 6

# The inhomogeneous universe: gravity

In the previous chapters, we took care of all non-gravitational interactions via the Boltzmann equations, and took into account the effects of gravity on the particle distributions. This formalism led to the set of Eqs. (5.67)–(5.73). We now need to supplement these equations with an account of how the perturbations to the particle distributions in turn affect the gravitational field. For this, we need the Einstein equations of general relativity. In Ch. 3, we derived the homogeneous solution of the Einstein equations. Here, we will expand perturbatively to linear order around the zeroth-order homogeneous solution. This calculation is completely straightforward. While a bit long, working through it is a "must-do-once" exercise so the steps are presented in some detail. First though, we will think about how to best break down the 10 independent Einstein equations, and how we choose our coordinates.

## 6.1 Scalar–vector–tensor decomposition

In the previous chapter, we have seen that the transformation to Fourier space simplified the perturbed Boltzmann equations considerably, by decoupling the different Fourier modes $\boldsymbol{k}$. The Einstein equations are a tensor equality and correspondingly comprise a set of equations that are in general coupled. However, there exists a decomposition of these equations that again allows us to decouple different modes. In fact, we have already implicitly relied on this result when writing the perturbed metric in the simple form of Eq. (3.49). We thus need to begin with this decomposition.

Let us consider an FLRW spacetime that is perturbed by a small amount. That is, we write the metric as

$$\begin{aligned} g_{00}(t,\boldsymbol{x}) &= -1 + h_{00}(t,\boldsymbol{x}), \\ g_{0i}(t,\boldsymbol{x}) &= a(t)h_{0i}(t,\boldsymbol{x}) = a(t)h_{i0}(t,\boldsymbol{x}), \\ g_{ij}(t,\boldsymbol{x}) &= a^2(t)\left[\delta_{ij} + h_{ij}(t,\boldsymbol{x})\right], \end{aligned} \tag{6.1}$$

where $h_{00}$, $h_{0i}$, $h_{ij}$ are metric perturbations that are functions of space and time, and all of whose components are assumed to be small in magnitude. In this chapter, we will mostly use the physical time $t$. Keep in mind that $t$ and the conformal time $\eta$ are always related through $dt = ad\eta$, so that going back and forth is a simple variable transformation that

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00012-7
Copyright © 2021 Elsevier Inc. All rights reserved.

does not touch the perturbations and is unrelated to the coordinate choice we will discuss below.

When we introduced the perturbed metric in Sect. 3.3, we assumed a special form of $h_{\mu\nu}$:

$$\begin{aligned} h_{00} &= -2\Psi, \\ h_{0i} &= 0, \\ h_{ij} &= 2\Phi\delta_{ij}\,, \end{aligned} \tag{6.2}$$

which we referred to as conformal-Newtonian gauge. Now let us take some time to study the perturbed FLRW metric more systematically. First, this will help us isolate the relevant components of the Einstein equations for the perturbed metric (which are the ultimate goal of this chapter). Second, we will see that the two potentials $\Phi$, $\Psi$ are not sufficient to capture all gravitational physics.

Technically, we want to classify the components of the general metric in Eq. (6.1) via their behavior under spatial rotations. The time-time component $h_{00}$ is a 3-scalar, i.e. it remains unchanged under spatial rotations, since it does not have a spatial index. To keep the discussion general, and not confined to conformal-Newtonian gauge, we will define this scalar $h_{00} = -2A$, with signs and factors of 2 here and in the following being a matter of convention. The time-space perturbation $h_{0i}$ is a 3-vector. So let us decompose it into longitudinal and transverse parts captured by two functions, $B$ and $B_i$:

$$h_{0i} = -\frac{\partial B}{\partial x^i} - B_i \quad \text{where} \quad {B_i}^{,i} \equiv \frac{\partial B_i}{\partial x^i} = 0. \tag{6.3}$$

Here, we use the comma notation to indicate ordinary partial (not covariant) derivatives with respect to coordinates, $B_{,i} \equiv \partial B/\partial x^i$. Notice again that here and throughout, we raise and lower spatial indices with $\delta_{ij}$, the appropriate spatial background metric for a Euclidean universe, and employ the sum convention over repeated indices. Thus, the first part of $h_{0i}$ is the *gradient of a 3-scalar* function $B(t,\boldsymbol{x})$, while the second is a divergenceless 3-vector $B_i(t,\boldsymbol{x})$. We will refer to the first as the "scalar," and the second as the "vector" contribution. In Fourier space, Eq. (6.3) becomes simpler:

$$h_{0i}(t,\boldsymbol{k}) = -ik_i B(t,\boldsymbol{k}) - B_i(t,\boldsymbol{k})\,, \quad \text{with} \quad k^i B_i = 0. \tag{6.4}$$

Next up is $h_{ij}$, which is a symmetric 3-tensor. Consider first its scalar contributions. We already know that one such contribution is simply proportional to the Kronecker delta, as written in Eq. (6.2), which we in general shall call $D$. But we can also generate a symmetric 3-tensor by taking two spatial derivatives of a scalar $E$. Similarly, we can construct a vector contribution to $h_{ij}$ by taking a derivative of a divergence-free vector field $V_i$. Thus, we have

$$h_{ij} = 2D\delta_{ij} - 2E_{,ij} + V_{i,j} + V_{j,i} \qquad \text{(scalar and vector)}, \tag{6.5}$$

where ${V_i}^{,i} = 0$, and we made sure that the vector contribution is symmetric in $i$ and $j$. Is this the complete decomposition of $h_{ij}$? The answer is no: there is another component of

$h_{ij}$ that cannot be written through derivatives acting on a scalar or a vector. One way to see this is to count degrees of freedom: $h_{\mu\nu}$ is a symmetric tensor in four dimensions, and thus contains 10 free functions (16 minus the six symmetry constraints on the off-diagonal components). So far, we have identified four scalar functions ($A$, $B$, $D$, $E$) and two transverse vectors ($B_i$, $V_i$); due to the transversality constraint, each vector contains two free functions. Thus we have so far enumerated eight functions in the metric. The two remaining ones are referred to as the *tensor* degrees of freedom, which we write as $h_{ij}^{\rm TT}$. We will see in Sect. 6.4 that they describe very important physics, namely propagating gravitational waves.

To summarize, our spatial metric perturbation becomes, in Fourier space,

$$h_{ij} = 2D\delta_{ij} + 2k_i k_j E + ik_i V_j + ik_j V_i + h_{ij}^{\rm TT}. \tag{6.6}$$

Here, we have concentrated on the decomposition of the metric. But there was nothing particular about $h_{\mu\nu}$, and we can do the exact same formal decomposition with any tensor, for example the energy-momentum tensor. A very important result in relativity, the *decomposition theorem*, states that perturbations of each type—scalar, vector, and tensor—evolve independently at linear order.[1] That is, if some physical process in the early universe sets up tensor perturbations, these do not induce scalar perturbations as they evolve. Conversely, to determine the evolution of scalar perturbations, we will not have to worry about possible vector or tensor perturbations. This is the justification for why we were able to neglect vector and tensor contributions in the perturbed metric up to now. The fundamental reason why the decomposition theorem holds is that the background FLRW metric is spatially isotropic. So, in going to Fourier space (which decouples different perturbations due to the homogeneity of the FLRW metric, see Box 5.1), and performing a scalar/vector/tensor decomposition, we have used the symmetries of the FLRW metric to our maximum advantage.

You probably noticed another difference between the metric perturbation in Eq. (6.2) and the general expressions we derived below it: even when restricting to scalar perturbations only, do we not need $B_{,i}$ and $E_{,ij}$ in $h_{0i}$ and $h_{ij}$, respectively? The reason we do not will become clear once we consider the effect of coordinate transformations.

## 6.2 From gauge to gauge

The effect of coordinate transformations on spacetime scalars, vectors, and tensors is described in Box 2.2. In the context of perturbation theory in relativity, a choice of coordinates is often referred to as *gauge*, and we will frequently use that term also. The ability to move back and forth between different gauges is useful when dealing with cosmological perturbations. Often, the equations simplify considerably in one gauge (one example is the *spa-*

[1] This decomposition refers to 3-scalars, 3-vectors, 3-tensors, not to be confused with spacetime scalars such as a field $\phi(x)$, spacetime vectors such as the momentum $P^\mu$, and tensors such as the metric $g_{\mu\nu}$. We should apologize for this confusing nomenclature, which, however, is widely used in the literature.

*tially flat slicing* when calculating perturbations generated during inflation, Sect. 7.4.3), while quantities that we actually measure observationally are more naturally calculated in another. So, different gauges have their advantages for different parts of the "cosmological perturbations" problem.

Let us start out with a scalar field $\phi(x)$, where in this section $x$ stands for a spacetime location $(t, \boldsymbol{x})$. We will encounter such a field when studying inflation in the next chapter. We are interested in small perturbations around the homogeneous universe, so we separate $\phi$ into background and perturbation:

$$\phi(x) = \bar{\phi}(t) + \delta\phi(t, \boldsymbol{x}), \tag{6.7}$$

where the background field can only depend on $t$ since the background universe is homogeneous. We now want to derive how Eq. (6.7) changes when we transform coordinates to $x \to \hat{x}(x)$. In keeping with our interest in perturbations, it is sufficient to consider small coordinate transformations as well; otherwise, the transformed field would in general have large (unphysical) perturbations. Hence, we perform a Taylor series of $\hat{x}(x)$ in $x$, and keep only the zeroth-order piece, which is a shift in coordinates. That is, we write

$$\begin{aligned} t \to \hat{t} &= t + \zeta(t, \boldsymbol{x}), \\ x^i \to \hat{x}^i &= x^i + \xi^{,i}(t, \boldsymbol{x}), \end{aligned} \tag{6.8}$$

where the time shift is $\zeta$ while the spatial coordinate shift is written as the gradient of another scalar function $\xi$, since we are considering only scalar perturbations for now (we will get back to this point below).

Treating $\delta\phi$, $\xi$, and $\zeta$ as first-order perturbations, the scalar transformation law (Box 2.2)

$$\hat{\phi}(\hat{x}) = \phi(x[\hat{x}]) = \phi(\hat{t} - \zeta, \hat{\boldsymbol{x}} - \boldsymbol{\nabla}\xi) \tag{6.9}$$

becomes (see Exercise 6.3)

$$\delta\hat{\phi}(t, \boldsymbol{x}) = \delta\phi(t, \boldsymbol{x}) - \frac{d\bar{\phi}(t)}{dt}\zeta(t, \boldsymbol{x}), \tag{6.10}$$

where we have dropped the hats on the coordinates for clarity. In other words, while the scalar $\phi(x)$ itself transforms trivially, i.e. $\hat{\phi}(\hat{x}) = \phi(x)$, its *perturbation* $\delta\phi$ does not. This is because, in order to define a perturbation, we need to assume a background value, which, if the field evolves in time at the background level, depends on the choice of time coordinate.

After this warmup, let us apply the small coordinate transformation Eq. (6.8) to the general perturbed metric with scalar perturbations only. Using the decomposition from the previous section, we can write this as

$$\begin{aligned} g_{00} &= -(1 + 2A), \\ g_{0i} &= -aB_{,i}, \end{aligned}$$

$$g_{ij} = a^2(\delta_{ij}[1+2D] - 2E_{,ij}). \tag{6.11}$$

As derived in Box 2.2 [Eq. (2.49)], the metric changes under a coordinate transformation $x \to \hat{x}(x)$ as

$$\hat{g}_{\mu\nu}(\hat{x}) = \frac{\partial x^\alpha}{\partial \hat{x}^\mu}\frac{\partial x^\beta}{\partial \hat{x}^\nu} g_{\alpha\beta}(x)\,, \tag{6.12}$$

or equivalently[2]

$$\hat{g}_{\alpha\beta}(\hat{x})\frac{\partial \hat{x}^\alpha}{\partial x^\mu}\frac{\partial \hat{x}^\beta}{\partial x^\nu} = g_{\mu\nu}(x). \tag{6.13}$$

We are now ready to work out the metric transformation explicitly. We will do so for one component and leave the rest as an exercise. Consider the $_{00}$ component of Eq. (6.13):

$$\hat{g}_{\alpha\beta}(\hat{x})\frac{\partial \hat{x}^\alpha}{\partial t}\frac{\partial \hat{x}^\beta}{\partial t} = -[1+2A]. \tag{6.14}$$

Note that, since $A$ is a perturbation, we do not need to distinguish between $x$ or $\hat{x}(x)$ in its argument, since the difference between the coordinates is itself a first-order quantity. Now, our claim is that the only term that contributes to the left-hand side is the one with $\alpha = \beta = 0$. Consider for example $\alpha = 0$ and $\beta = i$. The off-diagonal component $\hat{g}_{0i}$ of the metric is proportional to $\hat{B}_{,i}$ a first-order perturbation. But $\partial \hat{x}^i/\partial t$ is proportional to the first-order variable $\xi$, so the product is second-order and can be neglected. A similar argument holds for the $\alpha = i; \beta = j$ terms. Therefore, the left-hand side is simply

$$\begin{aligned} -[1+2\hat{A}]\left(\frac{\partial \hat{t}}{\partial t}\right)^2 &= -[1+2\hat{A}]\left(1+\dot{\xi}\right)^2 \\ &\simeq -1-2\hat{A}-2\dot{\xi}. \end{aligned} \tag{6.15}$$

Equating this with $g_{00}$ leads to

$$-2\hat{A} - 2\dot{\xi} = -2A, \tag{6.16}$$

so under the coordinate transformation specified by Eq. (6.8)

$$A \to \hat{A} = A - \frac{1}{a}\zeta'. \tag{6.17}$$

Similarly, the other components of the metric transform into

$$\begin{aligned} \hat{B} &= B - a^{-1}\zeta + \xi', \\ \hat{D} &= D - H\zeta, \\ \hat{E} &= E + \xi. \end{aligned} \tag{6.18}$$

[2] Any well-defined coordinate transformation has to have a nonzero determinant $|\partial x/\partial \hat{x}|$ so that this matrix is invertible.

You can prove these relations in Exercise 6.2. Eqs. (6.17)–(6.18) describe how a metric with small scalar perturbations transforms under a small scalar coordinate transformation. As expected from the decomposition theorem, this coordinate transformation did not generate any non-scalar metric perturbations, so that the result can still be written in terms of the functions $A$, $B$, $D$, $E$.

To sum up, then, there are four functions that characterize scalar metric perturbations, but these can be manipulated with two other functions that characterize scalar coordinate transformations. For example, starting with a metric in which $E \neq 0$, it is easy to make a transformation to eliminate $E$: simply choose $\xi = -E$, and the resulting metric has $\hat{E} = 0$. Thus, there are really only $4 - 2 = 2$ physical degrees of freedom describing scalar metric perturbations. Indeed, this is the reason we were able to eliminate $B$ and $E$ in conformal-Newtonian gauge. One can also take specific linear combinations of metric perturbations that are invariant under Eqs. (6.17)–(6.18); again there are exactly two independent such combinations. The most popular choice is that of Bardeen (1980):

$$\begin{aligned} \Phi_A &\equiv A + \frac{1}{a}\frac{\partial}{\partial\eta}[a(E' - B)], \\ \Phi_H &\equiv -D + aH(B - E'). \end{aligned} \tag{6.19}$$

In conformal-Newtonian gauge, in which $E = B = 0$, we have $\Phi_A = \Psi$ and $\Phi_H = -\Phi$. These invariants are very useful: if equations simplify in a particular gauge, then one can do calculations in that gauge, form the gauge-invariant variables, and then turn these into the perturbations in any other gauge. We will do precisely this in Sect. 7.4.3. In other words, $\Phi_A$ and $\Phi_H$ are useful shortcuts for transforming from one gauge to another.

Again, none of the mathematical derivations we just did are specific to the metric; Eq. (6.13) applies similarly to the energy-momentum tensor with two lower indices, $T_{\mu\nu}$. For matter in conformal-Newtonian gauge, the two scalar degrees of freedom that we have reduced perturbations in $T_{\mu\nu}$ to are the density perturbation $\delta_s$ and the longitudinal velocity $u_s$, for each species $s$.

Finally, going beyond scalar perturbations, we can already guess what will happen: we now have an additional degree of freedom in our coordinate transformation, a transverse vector $\xi^i$ in the spatial coordinate transformation. This will allow us to set one of the two vector metric perturbations $B_i$ or $V_i$ to zero, for example. We thus reduce the vectors from four to two independent degrees of freedom. And we may also guess, correctly, that small coordinate transformations of either the scalar or vector type leave the tensor perturbation $h_{ij}^{\mathrm{TT}}$ unchanged: it is gauge-invariant at linear order. Counting degrees of freedom, this again makes sense. We started with 10 degrees of freedom in the perturbed metric, but the coordinate transformation Eq. (6.8) allows us to remove four of them, so that we are left with six: two each of scalar, vector, and tensor types.

## 6.3 The Einstein equations for scalar perturbations

We are now ready to embark on our computation of the Einstein equations at linear order in perturbations. To begin, we will focus on scalar perturbations and continue to work in conformal-Newtonian gauge, so that our starting point is

$$\begin{aligned} g_{00}(\boldsymbol{x},t) &= -1-2\Psi(\boldsymbol{x},t), \\ g_{0i}(\boldsymbol{x},t) &= 0, \\ g_{ij}(\boldsymbol{x},t) &= a^2(t)\delta_{ij}[1+2\Phi(\boldsymbol{x},t)]. \end{aligned} \tag{6.20}$$

Evaluating the left-hand side of the Einstein equation (3.1) requires three steps:

- Compute the Christoffel symbol, $\Gamma^{\mu}{}_{\alpha\beta}$, for the perturbed metric of Eq. (6.20); we have already done this in Sect 3.3.1.
- From these, form the Ricci tensor, $R_{\mu\nu}$, using Eq. (3.3).
- Contract the Ricci tensor to form the Ricci scalar, $R \equiv g^{\mu\nu}R_{\mu\nu}$.

We will also immediately switch to Fourier space, exchanging spatial derivatives with powers of $i\boldsymbol{k}$. We need two independent equations for the two variables $\Phi$, $\Psi$. Given that we are dealing with scalar perturbations, we can already anticipate that the 00 component as well as the scalar component of the $ij$ Einstein equations will be useful.

### 6.3.1 Ricci tensor

The Ricci tensor is most easily expressed in terms of the Christoffel symbol we derived in Sect. 3.3.1. First, consider the time-time component of Eq. (3.3):

$$R_{00} = \Gamma^{\alpha}{}_{00,\alpha} - \Gamma^{\alpha}{}_{0\alpha,0} + \Gamma^{\alpha}{}_{\beta\alpha}\Gamma^{\beta}{}_{00} - \Gamma^{\alpha}{}_{\beta 0}\Gamma^{\beta}{}_{0\alpha}. \tag{6.21}$$

All of these terms contribute at first order. One simplification comes from considering the $\alpha = 0$ part of all these terms. The first and second terms are equal and opposite to each other as are the last two. So the sum over the index $\alpha$ contributes only when $\alpha$ is spatial. Let us consider each of the terms one by one.

- The first is

$$\Gamma^{i}{}_{00,i} = -\frac{k^2}{a^2}\Psi, \tag{6.22}$$

  after translating the first line of Eq. (3.56) into Fourier space.
- The second term in Eq. (6.21) is

$$-\Gamma^{i}{}_{0i,0} = -3\left(\frac{\ddot{a}}{a} - H^2 + \Phi_{,00}\right) \tag{6.23}$$

  using the second line of Eq. (3.56). The factor of 3 in front comes from the implicit sum in $\delta_{ii}$.

- The next term is $\Gamma^i{}_{i\beta}\Gamma^\beta{}_{00}$. Note that $\Gamma^\beta{}_{00}$ is first order no matter what $\beta$ is, so we need keep only the zeroth-order part of $\Gamma^i{}_{i\beta}$. However, the last line of Eq. (3.56) shows that $\Gamma^i{}_{i\beta}$ is first-order unless $\beta = 0$. So to first-order,

$$\begin{aligned}\Gamma^i{}_{i\beta}\Gamma^\beta{}_{00} &= \Gamma^i{}_{i0}\Gamma^0{}_{00} \\ &= 3H\Psi_{,0}.\end{aligned} \tag{6.24}$$

- Finally the last term is $-\Gamma^i{}_{\beta 0}\Gamma^\beta{}_{0i}$. In this case, if $\beta = 0$ both $\Gamma$ are first-order, so their product is second-order and can be neglected. Therefore, only spatial $\beta$ need to be considered, leading to

$$\begin{aligned}-\Gamma^i{}_{\beta 0}\Gamma^\beta{}_{0i} &= -\Gamma^i{}_{j0}\Gamma^j{}_{0i} \\ &= -3(H^2 + 2H\Phi_{,0}).\end{aligned} \tag{6.25}$$

Collecting these four sets of terms gives

$$R_{00} = -3\frac{\ddot{a}}{a} - \frac{k^2}{a^2}\Psi - 3\Phi_{,00} + 3H(\Psi_{,0} - 2\Phi_{,0}). \tag{6.26}$$

Note that the zeroth-order term agrees with Eq. (3.6). The space-space part of the Ricci tensor is left as an exercise. It is

$$\begin{aligned}R_{ij} = \delta_{ij}\Big[&\left(2a^2H^2 + a\ddot{a}\right)(1 + 2\Phi - 2\Psi) \\ &+ a^2H(6\Phi_{,0} - \Psi_{,0}) + a^2\Phi_{,00} + k^2\Phi\Big] + k_ik_j(\Phi + \Psi).\end{aligned} \tag{6.27}$$

We can now contract the indices on the Ricci tensor and find the Ricci scalar:

$$\begin{aligned}R \equiv g^{\mu\nu}R_{\mu\nu} &= g^{00}R_{00} + g^{ij}R_{ij} \\ &= [-1 + 2\Psi]\left[-3\frac{\ddot{a}}{a} - \frac{k^2}{a^2}\Psi - 3\Phi_{,00} + 3H(\Psi_{,0} - 2\Phi_{,0})\right] \\ &\quad + \frac{1 - 2\Phi}{a^2}\Bigg[3\Big\{\left(2a^2H^2 + a\ddot{a}\right)(1 + 2\Phi - 2\Psi) \\ &\qquad + a^2H(6\Phi_{,0} - \Psi_{,0}) + a^2\Phi_{,00} + k^2\Phi\Big\} + k^2(\Phi + \Psi)\Bigg].\end{aligned} \tag{6.28}$$

First let us check the zeroth-order part of $R$. Combining terms, we find that it is $6\left(H^2 + \ddot{a}/a\right)$, in agreement with Eq. (3.9). To get the first-order part, $\delta R$, we go through the by-now-familiar routine of multiplying terms, keeping only those first-order in $\Phi$ and $\Psi$. This gives

$$\begin{aligned}\delta R = &-6\Psi\frac{\ddot{a}}{a} + \frac{k^2}{a^2}\Psi + 3\Phi_{,00} - 3H(\Psi_{,0} - 2\Phi_{,0}) \\ &- 6\Psi\left(2H^2 + \frac{\ddot{a}}{a}\right) + 3H(6\Phi_{,0} - \Psi_{,0})\end{aligned}$$

$$+ 3\Phi_{,00} + 4\frac{k^2\Phi}{a^2} + \frac{k^2\Psi}{a^2}, \tag{6.29}$$

where the first line contains the terms from $R_{00}$ (the second line in Eq. (6.28)) and the last two lines come from $R_{ij}$. Combining these leads to

$$\begin{aligned}\delta R = &- 12\Psi\left(H^2 + \frac{\ddot{a}}{a}\right) + \frac{2k^2}{a^2}\Psi + 6\Phi_{,00} \\ &- 6H(\Psi_{,0} - 4\Phi_{,0}) + 4\frac{k^2\Phi}{a^2}.\end{aligned} \tag{6.30}$$

### 6.3.2 Two components of the Einstein equations

We can now derive the evolution equations for $\Phi$ and $\Psi$, our scalar perturbations to the Friedmann–Lemaître–Robertson–Walker metric. We have several different options here, because the Einstein equations

$$G^{\mu}{}_{\nu} = 8\pi G T^{\mu}{}_{\nu} \tag{6.31}$$

have 10 components and we need only two. All of the other eight components will either be zero at first-order or redundant.[3]

The first component we will use is the time-time component. Thus we need to evaluate

$$\begin{aligned}G^0{}_0 &= g^{00}\left[R_{00} - \frac{1}{2}g_{00}R\right] \\ &= (-1 + 2\Psi)R_{00} - \frac{R}{2}.\end{aligned} \tag{6.32}$$

Here one of the indices has been raised by multiplying $G_{00}$ by $g^{00}$ (recall that the $g^{0i}$ vanish). This turns out to simplify the energy-momentum tensor (see Sect. 3.4 and Exercise 3.12) which supplies the right-hand side. Also note that the second line follows from the first since $g^{00}g_{00} = 1$. We have computed the time-time component of the Ricci tensor (Eq. (6.26)) and the perturbed Ricci scalar (Eq. (6.30)), so the first-order part of the time-time component of the Einstein tensor is

$$\begin{aligned}\delta G^0{}_0 = &- 6\Psi\frac{\ddot{a}}{a} + \frac{k^2}{a^2}\Psi + 3\Phi_{,00} - 3H(\Psi_{,0} - 2\Phi_{,0}) \\ &+ 6\Psi\left(H^2 + \frac{\ddot{a}}{a}\right) - \frac{k^2}{a^2}\Psi - 3\Phi_{,00} \\ &+ 3H(\Psi_{,0} - 4\Phi_{,0}) - 2\frac{k^2\Phi}{a^2}.\end{aligned} \tag{6.33}$$

[3] This is true for scalar perturbations. When we come to consider tensor perturbations, some of the other components will be useful.

Combining terms leads to

$$\delta G^0{}_0 = -6H\Phi_{,0} + 6\Psi H^2 - 2\frac{k^2\Phi}{a^2}. \tag{6.34}$$

Einstein's equation equates $G^0{}_0$ with $8\pi G T^0{}_0$ where $T_{\mu\nu}$ is the energy-momentum tensor. To complete our derivation of the first evolution equation for $\Phi$ and $\Psi$, therefore, we need to compute the first-order part of the source term, $T^0{}_0$. Recall from Sect. 2.3 that $-T^0{}_0$ is the energy density of all the particles in the universe, and that the contribution from each species is an integral over the distribution function. In Ch. 3, we showed that even when including perturbations, the simple expression of Eq. (2.62) remains valid (i.e. Eq. (3.86)),

$$T^0{}_0(\boldsymbol{x},\ t) = -\sum_s g_s \int \frac{d^3p}{(2\pi)^3} E_s(p) f_s(\boldsymbol{p}, \boldsymbol{x}, t), \tag{6.35}$$

where the sum runs over species $s$, with degeneracy factor $g_s$, distribution function $f_s$, and energy-momentum relation $E_s(p) = \sqrt{p^2 + m_s^2}$. To get the first-order part of the energy-momentum tensor, we must naturally consider the first-order part of the distribution functions, i.e. the perturbation variables we defined in Ch. 5 for the photons, neutrinos, dark matter, and baryons.

This is easiest for the dark matter and baryons, which are non-relativistic so that $E_s(p) \simeq m_s$. Then their contribution to $T^0{}_0$ is simply proportional to $-m\,n(t, \boldsymbol{x})$, where $n$ is the number density of baryons and dark matter. We thus have

$$T^0{}_0\Big|_{s=\mathrm{b,c}} = -\rho_s(1+\delta_s). \tag{6.36}$$

For photons, a little more care is required. Using Eq. (5.3), we have

$$T^0{}_0\Big|_\gamma = -2\int \frac{d^3p}{(2\pi)^3} p\left[f^{(0)} - p\frac{\partial f^{(0)}}{\partial p}\Theta\right]. \tag{6.37}$$

The first term here is just the zeroth-order photon energy density, $\rho_\gamma$. To reduce the second term, we first do the angular integral, which picks out the monopole $\Theta_0$ from $\Theta$. Then, we do the integral over $p$ by parts. This changes the sign and introduces a factor of 4 since $\partial p^4/\partial p = 4p^3$, leading to

$$T^0{}_0\Big|_\gamma = -\rho_\gamma[1 + 4\Theta_0]. \tag{6.38}$$

The factor of 4 here is obvious in retrospect. The perturbation variable $\Theta$ is the fractional temperature change, while the energy-momentum tensor is interested in the perturbed energy density, $\delta\rho_\gamma$. We should have expected that, since $\rho_\gamma \propto T^4$, $\delta\rho_\gamma/\rho_\gamma = 4\delta T/T$. Be warned, however, that the literature is split between those who define $\Theta$ as $\delta\rho_\gamma/\rho_\gamma$ and those who opt for the convention we use here, which then differ by a factor of 4. Finally, the

first-order contribution from massless neutrinos is identical in form to the photon case,

$$T^0{}_0\Big|_{\nu,m_\nu=0} = -\rho_\nu[1+4\mathcal{N}_0]. \tag{6.39}$$

For neutrinos with mass, the integral over momentum can no longer be solved in closed form. So, for our analytic solutions in later chapters we will neglect neutrino masses, and we will discuss the impact of neutrino masses based on the numerical solution. Finally, we continue to neglect dark energy perturbations as discussed at the beginning of Ch. 5.

Returning to Einstein's equation, we equate Eq. (6.34) with $8\pi G$ times the first-order part of the time-time component of the energy-momentum tensor. Dividing both sides by 2 leads to

$$-3H\Phi_{,0} + 3\Psi H^2 - \frac{k^2\Phi}{a^2} = -4\pi G\left[\rho_{\rm c}\delta_{\rm c} + \rho_{\rm b}\delta_{\rm b} + 4\rho_\gamma\Theta_0 + 4\rho_\nu\mathcal{N}_0\right]. \tag{6.40}$$

It is again useful to write the equation in terms of conformal time. This introduces an extra factor of $1/a$ every time a time derivative appears, so

$$k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) = 4\pi G a^2[\rho_{\rm c}\delta_{\rm c} + \rho_{\rm b}\delta_{\rm b} + 4\rho_\gamma\Theta_0 + 4\rho_\nu\mathcal{N}_0]. \tag{6.41}$$

This is our first evolution equation for $\Phi$ and $\Psi$. In the limit of no expansion ($a =$ constant), Eq. (6.41) reduces to the ordinary Poisson equation for gravity (in Fourier space): the left-hand side is $-\nabla^2\Phi$ while the right-hand side is $4\pi G a^2\delta\rho$. The terms proportional to $a'$ account for expansion and are typically important for modes with physical wavelengths ($\sim a/k$) comparable to, or larger than, the Hubble radius, $H^{-1}$. We need this general-relativistic expression when we consider the evolution of perturbations, because all modes of interest today once had wavelengths larger than the Hubble radius. More on this in Ch. 7.

We now need a second evolution equation for $\Phi$ and $\Psi$. Let us focus on the spatial part of $G^\mu{}_\nu$,

$$G^i{}_j = g^{ik}\left[R_{kj} - \frac{g_{kj}}{2}R\right] = \frac{\delta^{ik}(1-2\Phi)}{a^2}R_{kj} - \frac{\delta^i{}_j}{2}R. \tag{6.42}$$

From Eq. (6.27), we see that most of the terms in $R_{kj}$ are proportional to $\delta_{kj}$. When contracted with $\delta^{ik}$ this will lead to a host of terms proportional to $\delta_{ij}$, in addition to the last term here, the one proportional to $R$. Therefore, Eq. (6.42) can be written as

$$G^i{}_j = F(\Phi,\Psi)\delta^i{}_j + \frac{k^i k_j(\Phi+\Psi)}{a^2} \tag{6.43}$$

where $F(\Phi,\Psi)$ has close to a dozen terms which we would rather not write down. Since all of these terms are proportional to $\delta^i{}_j$ they all contribute to the trace of $G^i{}_j$. To avoid dealing with these terms, consider the *longitudinal, traceless* part of $G^i{}_j$, which can be extracted by contracting $G^i{}_j$ with $\hat{k}_i\hat{k}^j - (1/3)\delta^j_i$. This procedure picks out the piece which

is longitudinal and traceless, and only that part (Exercise 6.1). We are left with

$$\left(\hat{k}_i\hat{k}^j - \frac{1}{3}\delta_i^j\right) G^i{}_j = \left(\hat{k}_i\hat{k}^j - \frac{1}{3}\delta_i^j\right)\left(\frac{k^i k_j(\Phi+\Psi)}{a^2}\right) = \frac{2}{3a^2}k^2(\Phi+\Psi). \tag{6.44}$$

This is to be equated with the longitudinal, traceless part of the energy-momentum tensor, extracted in the same fashion. From Sect. 3.4, we have

$$T^i{}_j(\boldsymbol{x},t) = \sum_s g_s \int \frac{d^3p}{(2\pi)^3}\frac{p^i p_j}{E_s(p)} f_s(\boldsymbol{x},\boldsymbol{p},t). \tag{6.45}$$

Acting on this with the projection operator yields

$$\left(\hat{k}_i\hat{k}^j - \frac{1}{3}\delta_i^j\right) T^i{}_j = \sum_s g_s \int \frac{d^3p}{(2\pi)^3}\frac{p^2\mu^2 - (1/3)p^2}{E_s(p)} f_s(\boldsymbol{p}), \tag{6.46}$$

using the definition for $\mu$ via $\hat{\boldsymbol{k}}\cdot\boldsymbol{p} = \mu p$. We can immediately recognize the combination $\mu^2 - 1/3$ as proportional to the second Legendre polynomial, more precisely equal to $(2/3)\mathcal{P}_2(\mu)$. Therefore, the integral picks out the quadrupole part of the distribution. Since the zeroth-order part of the distribution function has no quadrupole, the source term is first order and nonzero only for photons and neutrinos, i.e. it is proportional to $\Theta_2$ and $\mathcal{N}_2$. The integral in Eq. (6.46) for photons is

$$\begin{aligned} -2\int \frac{dpp^2}{2\pi^2}p^2\frac{\partial f^{(0)}}{\partial p}\int_{-1}^{1}\frac{d\mu}{2}\frac{2\mathcal{P}_2(\mu)}{3}\Theta(\mu) &= 2\frac{2\Theta_2}{3}\int \frac{dpp^2}{2\pi^2}p^2\frac{\partial f^{(0)}}{\partial p} \\ &= -\frac{8}{3}\rho_\gamma\Theta_2 \end{aligned} \tag{6.47}$$

where the first equality follows from the definition of the quadrupole and the second from an integration by parts. This component of the energy-momentum tensor is called the *anisotropic stress*. Nonrelativistic particles, such as baryons and dark matter, do not contribute to the anisotropic stress, as the factor of $p/E_s(p)$ in Eq. (6.45) strongly suppresses their contribution.

For the second Einstein equation, we therefore equate Eq. (6.44) with $8\pi G$ times the photon and (massless) neutrino anisotropic stresses:

$$k^2(\Phi+\Psi) = -32\pi G a^2[\rho_\gamma\Theta_2 + \rho_\nu\mathcal{N}_2]. \tag{6.48}$$

This is an extremely important and useful result: it says that the two gravitational potentials are equal and opposite unless the photons or neutrinos have appreciable quadrupole moments. In practice, the photon quadrupole contributes little to this sum, because it is very small during the time when the photons have appreciable energy density (due to tight coupling; recall the argument after Eq. (5.22)). Only the collisionless neutrinos have an appreciable quadrupole moment early on when radiation dominates the universe.

Eq. (6.41) and Eq. (6.48) are the desired two Einstein equations for the metric perturbations $\Phi$, $\Psi$. A note to aficionados of differential equations: both equations do not contain any second time derivatives acting on $\Phi$, $\Psi$: they are *constraint equations*. That is, $\Phi$, $\Psi$ do not represent propagating degrees of freedom (neither does the familiar gravitational potential in Newtonian theory). This is a key difference from the tensor modes we turn to next.

## 6.4 Tensor perturbations

Until now, we have derived equations applying to the *scalar* perturbations of the homogeneous FLRW universe. This focus is reasonable: as we have seen, scalar perturbations to the metric are sourced by density fluctuations and vice versa. For the most part, the density fluctuations that form the structure of the universe are our primary interest. Moreover, thanks to the decomposition theorem it is perfectly fine to study scalar perturbations in isolation.

Nonetheless, we have seen in Sect. 6.1 that there are other types of gravitational perturbations, in particular *tensor* perturbations. In the next chapter we will see that the leading theory for the origin of scalar perturbations—inflation—also predicts tensor perturbations. Independently of cosmology, though, *gravitational waves* have emerged as a powerful probe of diverse astrophysical phenomena in the aftermath of their first detection by the LIGO collaboration. The wavelengths that LIGO is sensitive to are of order hundreds of kilometers, while we will be considering wavelengths of thousands of Mpc. However, the fundamental equation that governs their production and propagation is identical and we are now all set to derive that equation.

The most promising way to search for cosmological gravitational waves is through the distortions they induce in the CMB, especially on large scales. Sprinkled throughout the book, therefore, are exercises relating to tensor perturbations. The third type, vector perturbations, are also covered in the exercises. They are less interesting, since they are not sourced in appreciable amounts in most cosmological scenarios and, in any case, decay rapidly after they are produced. The tools needed to study vector and tensor modes are precisely those we crafted when studying scalar perturbations.

Tensor perturbations can be characterized by a metric perturbation (see Eq. (6.1)) with $h_{00} = -1$, $h_{0i} = 0$, and

$$\delta g_{ij}(t, \boldsymbol{x}) = a^2(t) h_{ij}^{\mathrm{TT}}(t, \boldsymbol{x}) \,, \quad h_{ij}^{\mathrm{TT}} = \begin{pmatrix} h_+ & h_\times & 0 \\ h_\times & -h_+ & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{6.49}$$

That is, the perturbations to the metric are described by two functions, $h_+$ and $h_\times$, assumed small. For definiteness, we have chosen the perturbations to be in the $x$–$y$ plane. This corresponds to an implicit choice of axes; in particular, it corresponds to choosing the $z$-axis to be in the direction of the wavevector, $\boldsymbol{k}$. More generally, $h_+$ and $h_\times$ are two components of

a *divergenceless, traceless, symmetric tensor*. *Divergenceless* means that $k^i h_{ij}^{\rm TT} = k^j h_{ij}^{\rm TT} = 0$.[4] This is clearly satisfied by Eq. (6.49) since there are no components in the $\hat{\boldsymbol{k}} = \hat{\boldsymbol{e}}_z$ direction. Tracelessness is also satisfied since the sum of the perturbations along the diagonal vanishes. For most of the derivation, we will only rely on the transverse and traceless nature of $h_{ij}^{\rm TT}$, and specialize to the case $\hat{\boldsymbol{k}} = \hat{\boldsymbol{e}}_z$ only at the very end.

Beyond the math, we will see shortly that the tensor-mode contribution to the metric Eq. (6.49) has distinctive geometric features. First though, with the metric written down, we can blast away and derive the Einstein equations. Once again, the derivation proceeds in two steps: (*i*) Christoffel symbol, and (*ii*) Ricci and Einstein tensors.

### 6.4.1 Christoffel symbol for tensor perturbations

First consider $\Gamma^0{}_{\alpha\beta}$. The metric we are considering in Eq. (6.49) has constant $g_{00}$ and vanishing $g_{0i}$. Recall that the Christoffel symbol is a sum of derivatives of the metric. The only terms that will be nonzero are those that involve derivatives of the spatial part of the metric, $g_{ij,\alpha}$. Therefore, we can immediately argue that

$$\Gamma^0{}_{00} = \Gamma^0{}_{i0} = 0. \tag{6.50}$$

The term with two lower spatial indices is

$$\Gamma^0{}_{ij} = -\frac{g^{00}}{2} g_{ij,0} = \frac{1}{2} g_{ij,0}. \tag{6.51}$$

Since $g_{ij} = a^2(\delta_{ij} + h_{ij}^{\rm TT})$, we have

$$g_{ij,0} = 2H g_{ij} + a^2 h_{ij,0}^{\rm TT}. \tag{6.52}$$

The first nonzero Christoffel symbol is therefore

$$\Gamma^0{}_{ij} = H g_{ij} + \frac{a^2 h_{ij,0}^{\rm TT}}{2}. \tag{6.53}$$

When both lower indices on $\Gamma$ are 0, the Christoffel symbol vanishes. The two remaining components are $\Gamma^i{}_{0j}$ and $\Gamma^i{}_{jk}$. The former is

$$\Gamma^i{}_{0j} = \frac{g^{ik}}{2} g_{jk,0}. \tag{6.54}$$

The time derivative of $g_{jk}$ acts on both the scale factor and on the perturbations $h_{+,\times}$, as in Eq. (6.52), so

$$\Gamma^i{}_{0j} = \frac{g^{ik}}{2}\left[2H g_{jk} + a^2 h_{jk,0}^{\rm TT}\right]. \tag{6.55}$$

[4] The divergence parts of $h_{ij}$ are contained in the scalar and vector perturbations $D$, $E$, $V_i$ in Eq. (6.6).

But $g^{ik}g_{jk} = \delta_{ij}$, so the first term here is simply the background contribution, $\delta_{ij}H$. For the second, we can set $g^{jk} = \delta_{jk}/a^2$ (i.e., neglect first-order terms) since it multiplies the first-order $h^{\rm TT}_{jk}$. So,

$$\Gamma^{i}{}_{0j} = H\delta_{ij} + \frac{1}{2}h^{\rm TT}_{ij,0}, \tag{6.56}$$

where we have used the fact that $h^{\rm TT}_{ij}$ is symmetric.

The last Christoffel symbol we need is $\Gamma^{i}{}_{jk}$. In Exercise 6.8 you will show that

$$\Gamma^{i}{}_{jk} = \frac{i}{2}\left[k_k h^{\rm TT}_{ij} + k_j h^{\rm TT}_{ik} - k_i h^{\rm TT}_{jk}\right]. \tag{6.57}$$

### 6.4.2 Ricci tensor for tensor perturbations

Following the same steps as in the scalar perturbation case, we now combine these Christoffel symbols to form the Ricci tensor. First we compute the time-time component $R_{00}$ of the Ricci tensor. Actually, we do not have to compute it explicitly: since $R_{00}$ has no spatial index (it is a *3-scalar*), we know that the indices of $h^{\rm TT}_{ij}$ have to be contracted with other indices inside $R_{00}$. Our only options are $\delta^{kl}$ and $k^i$; the indices could also be contracted with another factor of $h^{\rm TT}_{kl}$, but that would result in a second-order term. Now, since $h^{\rm TT}_{ij}$ is trace-free and divergenceless, all contractions with the Kronecker delta or $k^i$ vanish. This means that $R_{00}$ cannot contain a tensor-mode contribution at linear order; this is a manifestation of the decomposition theorem. In fact, the same holds for the Ricci scalar $R$.

The spatial components of the Ricci tensor do depend on the tensor perturbation variables. We have

$$R_{ij} = \Gamma^{\alpha}{}_{ij,\alpha} - \Gamma^{\alpha}{}_{i\alpha,j} + \Gamma^{\alpha}{}_{\alpha\beta}\Gamma^{\beta}{}_{ij} - \Gamma^{\alpha}{}_{\beta j}\Gamma^{\beta}{}_{i\alpha}. \tag{6.58}$$

Let us consider the first two terms together. Expanding out leads to

$$\Gamma^{\alpha}{}_{ij,\alpha} - \Gamma^{\alpha}{}_{i\alpha,j} = \Gamma^{0}{}_{ij,0} + \Gamma^{k}{}_{ij,k} - \Gamma^{k}{}_{ik,j} \tag{6.59}$$

since $\alpha = 0$ does not contribute in $\Gamma^{\alpha}{}_{i\alpha,j}$ because of Eq. (6.50). The lengthiest term here is the first, which involves multiple time derivatives. Let us postpone its calculation by recalling that $\Gamma^{0}{}_{ij} = g_{ij,0}/2$ so that the first term can be written in shorthand as $g_{ij,00}/2$. The last term in Eq. (6.59) vanishes since $\Gamma^{k}{}_{ik} = 0$ for tensor perturbations. Combining the other terms then leads to

$$\Gamma^{\alpha}{}_{ij,\alpha} - \Gamma^{\alpha}{}_{i\alpha,j} = \frac{g_{ij,00}}{2} + \frac{1}{2}\left[-k_i k_k h^{\rm TT}_{jk} - k_j k_k h^{\rm TT}_{ik} + k^2 h^{\rm TT}_{ij}\right]. \tag{6.60}$$

The first two terms in brackets vanish due to the transverse nature of $h^{\rm TT}_{ij}$. Therefore,

$$\Gamma^{\alpha}{}_{ij,\alpha} - \Gamma^{\alpha}{}_{i\alpha,j} = \frac{g_{ij,00}}{2} + \frac{k^2}{2}h^{\rm TT}_{ij}. \tag{6.61}$$

The third term in Eq. (6.58), $\Gamma^{\alpha}{}_{\alpha\beta}\Gamma^{\beta}{}_{ij}$, is nonzero only when the index $\alpha$ is spatial, so

$$\Gamma^{\alpha}{}_{\alpha\beta}\Gamma^{\beta}{}_{ij} = \Gamma^{k}{}_{k0}\Gamma^{0}{}_{ij} + \Gamma^{k}{}_{kl}\Gamma^{l}{}_{ij}. \tag{6.62}$$

But each of the Christoffel symbol in the second term here are of first order, so their product vanishes. In the first term, the sum over $k$ makes the first-order terms go away, so $\Gamma^{k}{}_{k0}$ is purely of zeroth order, $3H$. Therefore,

$$\Gamma^{\alpha}{}_{\alpha\beta}\Gamma^{\beta}{}_{ij} = \frac{3}{2}Hg_{ij,0}. \tag{6.63}$$

The final term in Eq. (6.58) will be left as an exercise; it is

$$\Gamma^{\alpha}{}_{\beta j}\Gamma^{\beta}{}_{i\alpha} = 2H^2 g_{ij} + 2a^2 H h^{\rm TT}_{ij,0}. \tag{6.64}$$

We can now combine all four terms in Eq. (6.58) to get

$$R_{ij} = \frac{g_{ij,00}}{2} + \frac{k^2}{2}h^{\rm TT}_{ij} + \frac{3}{2}Hg_{ij,0} - 2H^2 g_{ij} - 2a^2 H h^{\rm TT}_{ij,0}\,. \tag{6.65}$$

We now need to expand out the time derivatives of the metric. Using Eq. (6.52), one finds

$$g_{ij,00} = 2g_{ij}\left(\frac{\ddot{a}}{a} + H^2\right) + 4a^2 H h^{\rm TT}_{ij,0} + a^2 h^{\rm TT}_{ij,00}. \tag{6.66}$$

Therefore the Ricci tensor is

$$R_{ij} = g_{ij}\left(\frac{\ddot{a}}{a} + 2H^2\right) + \frac{3}{2}a^2 H h^{\rm TT}_{ij,0} + a^2\frac{h^{\rm TT}_{ij,00}}{2} + \frac{k^2}{2}h^{\rm TT}_{ij}. \tag{6.67}$$

Again we see that we have successfully recaptured the zeroth-order part of the Ricci tensor. Above, we argued that the Ricci scalar,

$$R = g^{00}R_{00} + g^{ij}R_{ij} \tag{6.68}$$

does not receive any contribution from $h^{\rm TT}_{ij}$ at linear order. Now, using Eq. (6.67), you can easily convince yourself of this fact. We are ready then to move on to the Einstein equations. As expected from Sect. 6.1, we will see that the first-order parts of the Einstein tensor we just computed do not couple to the scalar perturbations.

### 6.4.3 Einstein equations for tensor perturbations

Now let us read off the perturbations to the Einstein tensor induced by tensor modes. Since the Ricci scalar is unperturbed by tensor perturbations, the first-order Einstein tensor is simply

$$\delta G^{i}{}_{j} = \delta R^{i}{}_{j}. \tag{6.69}$$

To get $R^i{}_j$, we contract $g^{ik}R_{kj}$, using the Ricci tensor we computed in Eq. (6.67). The first term, proportional to the contraction of $g^{ik}g_{kj} = \delta^i{}_j$, has no first-order piece; the remaining terms are explicitly of first order in $h^{\rm TT}$, so we can set $g^{ik} = \delta^{ik}/a^2$, leading to

$$\delta G^i{}_j = \delta^{ik}\left[\frac{3}{2}Hh^{\rm TT}_{kj,0} + \frac{h^{\rm TT}_{kj,00}}{2} + \frac{k^2}{2a^2}h^{\rm TT}_{kj}\right]. \tag{6.70}$$

Finally, we specialize to the case of $\hat{\boldsymbol{k}} = \hat{\boldsymbol{e}}_z$ to derive a set of evolution equations for the tensor variables, $h_+$ and $h_\times$ (the final equation will be independent of this convenience choice).

To derive an equation for $h_+$, let us consider the difference between the ${}^1{}_1$ and ${}^2{}_2$ components of the Einstein tensor. The Einstein tensor in Eq. (6.70) is proportional to $h^{\rm TT}_{ij}$ and its derivatives. Since $h^{\rm TT}_{11} = -h^{\rm TT}_{22} = h_+$, $\delta G^1{}_1$ is equal and opposite to $\delta G^2{}_2$. Therefore,

$$\delta G_1{}^1 - \delta G_2{}^2 = 3Hh_{+,0} + h_{+,00} + \frac{k^2h_+}{a^2}. \tag{6.71}$$

Now we change to conformal time so that $h_{+,0} = h'_+/a$ and $h_{+,00} = h''_+/a^2 - (a'/a^3)h'_+$. Then,

$$a^2[\delta G^1{}_1 - \delta G^2{}_2] = h''_+ + 2\frac{a'}{a}h'_+ + k^2h_+. \tag{6.72}$$

The right-hand side of this component of Einstein's equations is zero in the absence of anisotropic stress (Exercise 6.9). This means that gravitational waves are not produced by the perturbations to matter that we derived in Ch. 5. Anisotropies in the radiation components (photons and neutrinos) do have an anisotropic stress, given by their quadrupole. As we argued in the previous section, for photons the quadrupole is suppressed during the radiation-dominated era, so their source term can be ignored. The most relevant quantity on the right-hand side of the tensor Einstein equations then is the neutrino anisotropic stress. This does provide a source term for gravitational waves, which leads to a damping of tensor modes on small scales. We neglect it here since we will focus on large-scale tensor modes throughout the rest of the book.

Finally, $h_\times$ obeys the same equation as $h_+$ (Exercise 6.11), so the tensor modes are governed by

$$h''_t + 2\frac{a'}{a}h'_t + k^2h_t = 0 \tag{6.73}$$

where $t = +, \times$. Eq. (6.73) is a wave equation, and the corresponding solutions are called *gravitational waves*. For example, if we neglect the expansion of the universe so that the damping term in Eq. (6.73) vanishes, we immediately see that the two solutions are $h_t \propto e^{\pm ik\eta}$. In real space, then, the perturbation to the metric is of the form

$$h_t(\boldsymbol{x},\eta) = \int \frac{d^3k}{(2\pi)^3}e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\left[A(\boldsymbol{k})e^{ik\eta} + B(\boldsymbol{k})e^{-ik\eta}\right] \qquad \text{(no expansion)}. \tag{6.74}$$

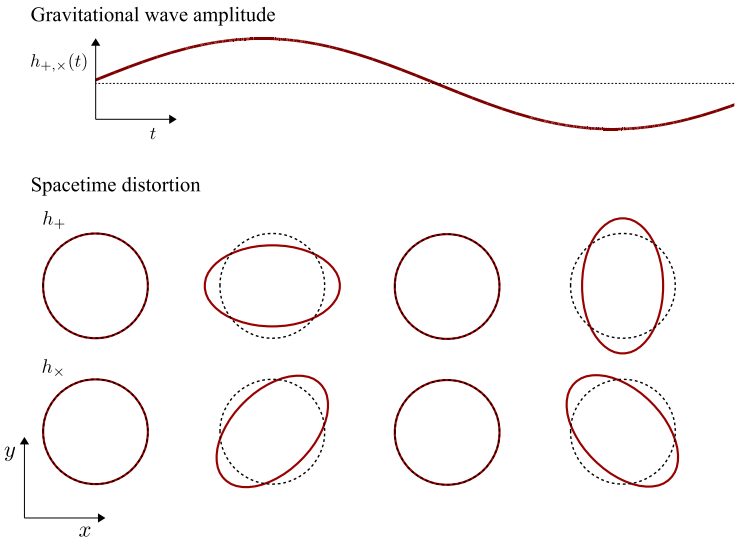



**FIGURE 6.1** Illustration of the perturbed spacetime due to a propagating gravitational wave (tensor mode); the $\boldsymbol{k}$-vector is along the $z$ axis as in Eq. (6.49), which comes out of the page. The upper panel shows the time evolution of the wave amplitude (neglecting the damping due to the expansion within one wave period), while the lower panel shows the stretching and compression of spacetime perpendicular to the wave's direction of propagation at the different points in time during a single wave cycle.

The two modes here correspond to waves traveling in the $\pm z$ direction at the speed of light. The spacetime distortions induced by these modes are illustrated in the lower panel of Fig. 6.1. Notice the elliptical pattern in the plane perpendicular to the wavevector (lower panel). This asymmetry is a fundamental distinction between tensor modes and scalar perturbations (see also Exercise 6.14); the spacetime distortion induced by scalar perturbations is always azimuthally symmetric around the wavevector.

Eq. (6.73) is the generalization of the gravitational-wave equation to an expanding universe. Exercise 6.12 shows that if the universe is purely radiation or matter dominated, exact analytic solutions can be obtained. These are oscillatory, like the simple ones in Eq. (6.74), but also decay in amplitude. Fig. 6.2 shows the evolution of $h_t$ for four modes of different wavelengths. Each mode remains constant at early times when its wavelength is larger than the horizon $k\eta < 1$. We will see in the next chapter what the term "larger than the horizon" means precisely; for now, simply notice that if we set $k \to 0$ in Eq. (6.73), $h_t =$ constant is a solution. Once the wavelength of the mode becomes comparable to the horizon, its amplitude oscillates with a frequency $k/2\pi$ and begins to decay. In particular, the decay is such ($\propto 1/a$) that the energy density in gravitational waves redshifts as $a^{-4}$, exactly like we expect for any form of radiation. Modes with a given $k$ are said to *enter the horizon* when $k\eta = 1$. Since the horizon entry of small-scale modes (e.g., $k/H_0 = 1000$ shown in Fig. 6.2) happens earlier, they have decayed more than large-scale tensor modes.

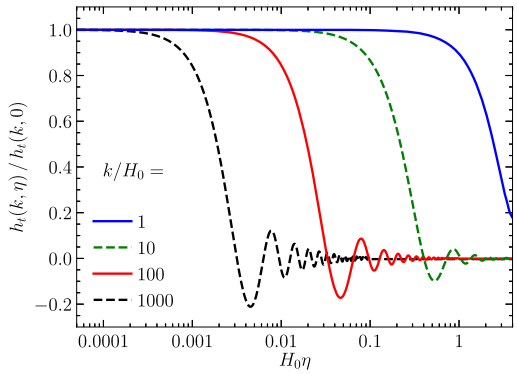



**FIGURE 6.2** Evolution of gravitational waves with different wavenumbers as a function of conformal time, normalized to their initial super-horizon amplitude. Each mode begins to oscillate and decay as its wavelength becomes smaller than the horizon, which corresponds to the epoch $k\eta = 1$. Smaller-scale (higher $k$) modes decay earlier.

### 6.4.4 Verifying the decomposition theorem

Now that we have computed the contributions to the Einstein tensor $G_{\mu\nu}$ from scalars and tensors, we can demonstrate the decomposition of these two types of perturbations. To do this, remember that we obtained the scalar equations by considering the two components of the Einstein tensor:

$$G^0{}_0; \qquad \left(\hat{k}_i\hat{k}_j - \frac{1}{3}\delta_{ij}\right) G^i{}_j. \tag{6.75}$$

Inserting these components into Einstein's equations led to Eq. (6.41) and Eq. (6.48). If we can show that tensor perturbations do not contribute to these two components, then we will have convinced ourselves of at least part of the decomposition theorem, namely that the equations governing scalar perturbations are not affected by tensors.

Tensor perturbations do not contribute to $G^0{}_0$, for $G^0{}_0$ depends on $R_{00}$ and $R$, and we have seen that both of these do not depend on $h_+$ or $h_\times$. Now let us show that $(\hat{k}_i\hat{k}_j - \delta_{ij}/3)G^i{}_j$ also does not pick up a contribution from tensor perturbations. Multiply Eq. (6.70) by the projection operator:

$$\left(\hat{k}_i\hat{k}_j - \frac{1}{3}\delta_{ij}\right)\delta G^i{}_j = \left(\hat{k}^i\hat{k}^j - \frac{1}{3}\delta^{ij}\right) \times \left[\frac{3}{2}Hh^{\rm TT}_{ij,0} + \frac{h^{\rm TT}_{ij,00}}{2} + \frac{k^2}{2a^2}h^{\rm TT}_{ij}\right]. \tag{6.76}$$

All the terms on the right-hand side are zero: they either involve contractions such as $\hat{k}^i h^{\rm TT}_{ij}$ (and time derivatives thereof), which vanish thanks to transversality, or the trace of $h^{\rm TT}_{ij}$, which vanishes since $h^{\rm TT}_{ij}$ is trace-free. The scalar equations we derived in the previous section are therefore unchanged by the presence of tensor modes. This is a manifestation of the decomposition theorem.

## 6.5 Summary

The Einstein equations relate perturbations in the metric to perturbations in the matter and radiation. We can decompose perturbations according to their behavior under spatial rotations: the scalar–vector–tensor–decomposition. Scalar, vector, and tensor perturbations are decoupled: each evolves independently of the others at linear order. Taking two components of the Einstein equations, we found equations for the evolution of the two functions that describe scalar metric perturbations, $\Phi$ and $\Psi$. As for the Boltzmann equations of the previous chapter, it is easiest to write these equations in Fourier space. Recalling our convention of dropping the tilde on Fourier-transformed variables (Eq. (5.29)), we can write

$$k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) = 4\pi G a^2[\rho_{\rm m}\delta_{\rm m} + 4\rho_{\rm r}\Theta_{\rm r,0}] \tag{6.77}$$

$$k^2(\Phi + \Psi) = -32\pi G a^2 \rho_{\rm r}\Theta_{\rm r,2}. \tag{6.78}$$

Here, the subscript m includes all matter, such as baryons and dark matter, and the subscript r all radiation, such as neutrinos and photons. More precisely,

$$\begin{aligned} &\rho_{\rm m}\delta_{\rm m} \equiv \rho_{\rm c}\delta_{\rm c} + \rho_{\rm b}\delta_{\rm b}; \qquad && \rho_{\rm r}\Theta_{\rm r,0} \equiv \rho_\gamma\Theta_0 + \rho_\nu\mathcal{N}_0; \\ &\rho_{\rm m}u_{\rm m} \equiv \rho_{\rm c}u_{\rm c} + \rho_{\rm b}u_{\rm b}; && \rho_{\rm r}\Theta_{\rm r,1} \equiv \rho_\gamma\Theta_1 + \rho_\nu\mathcal{N}_1. \end{aligned} \tag{6.79}$$

The other components of the Einstein equations are either zero or redundant; they add no new information about the evolution of $\Phi$ and $\Psi$ (this is similar to the case of the homogeneous universe, where all Einstein equations led to the Friedmann equations). An example of redundancy is the time-space component, which you can derive in Exercise 6.6. At times, though, one form of the evolution equation will be more useful than another. For example, one combination (Exercise 6.7) of these equations leads to an algebraic equation for the potential,

$$k^2\Phi = 4\pi G a^2\left[\rho_{\rm m}\delta_{\rm m} + 4\rho_{\rm r}\Theta_{\rm r,0} + \frac{3aH}{k}(i\rho_{\rm m}u_{\rm m} + 4\rho_{\rm r}\Theta_{\rm r,1})\right]. \tag{6.80}$$

Other components of Einstein's equation contain information not about the scalar perturbations $\Phi$ and $\Psi$, but about vector and tensor perturbations. While vector modes decay rapidly if they are not sourced, tensor modes are important as they describe gravitational waves. We will see in Ch. 7 that inflation also produces tensor perturbations, so it is important to know what the Einstein equations say about their evolution. We showed that there are two functions that characterize tensor perturbations, $h_+$ and $h_\times$; each of these evolves independently and satisfies

$$h_t'' + 2\frac{a'}{a}h_t' + k^2 h_t = 0 \tag{6.81}$$

where $t$ stands for $+$, $\times$. In an expanding universe, the amplitude of a gravitational wave described by Eq. (6.81) falls off once the mode enters the horizon.

There is excellent literature that treats the issues of gauge choices and the decomposi-

tion theorem in much more detail than we have here. *Cosmological Inflation and Large Scale Structure* (Liddle and Lyth, 2000) has a very nice treatment which, among other virtues, explains the physics of gauge choices. Two excellent, classic review articles on these topics are Mukhanov et al. (1992) and Kodama and Sasaki (1984). Finally, the seminal Bardeen (1980) article on gauge-invariant variables is remarkable for its clarity and conciseness.

## Exercises

**6.1** Consider a $3 \times 3$ tensor with Fourier-space components $G_{ij}(\boldsymbol{k}) = (\hat{k}_i \hat{k}_j - \delta_{ij}/3) G^L(\boldsymbol{k})$. Show that this form is traceless and satisfies $\epsilon_{ijk} G_{kl,jl} = 0$ in real space, where $\epsilon_{ijk}$ is the Levi-Civita symbol, so it is the proper form for the longitudinal component.

**6.2** Derive the transformations in the metric components given by Eq. (6.18). Show that $\Phi_A$ and $\Phi_H$ do not change under a general coordinate transformation.

**6.3** Derive Eq. (6.10) at linear order in perturbations. Use the fact that both $\delta\phi$ and the shift vector $\xi^\mu$ are first order in perturbations.

**6.4** Derive the Christoffel symbol, $\Gamma^i{}_{\mu\nu}$, given in Eq. (3.56). When doing this, you will need $g^{ij}$; show that $g^{ij} = \delta^{ij}(1 - 2\Phi)/a^2$.

**6.5** Show that $R_{ij}$ is given by Eq. (6.27).

**6.6** Compute the time-space component of the Einstein equations. Show that, in Fourier space, the relevant component of the Einstein tensor is

$$G^0{}_i = 2ik_i \left( \frac{\Phi'}{a} - H\Psi \right). \tag{6.82}$$

Combine with the energy-momentum tensor given in Eq. (3.86) to show that

$$\Phi' - aH\Psi = \frac{4\pi G a^2}{ik} [\rho_c u_c + \rho_b u_b - 4i\rho_\gamma \Theta_1 - 4i\rho_\nu \mathcal{N}_1]. \tag{6.83}$$

Notice that the integral over the distribution function in $T^0{}_i$ is already of first order, so you can neglect $\Phi$, $\Psi$ in the prefactor in Eq. (3.86). The time-space component of Einstein's equations adds no new information once we already have the two equations derived in the text. Deciding which two to use is a matter of convenience.

**6.7** Combine the time-time equation (6.41) with the time-space equation of Exercise 6.6 to obtain the algebraic (i.e., no time derivatives) equation for the potential given in Eq. (6.80). Show that this reduces to Poisson's equation, with the appropriate factors of $a$, when the wavelength of the perturbation is much smaller than the horizon ($k/aH \gg 1$), i.e. in the "Newtonian limit."

**6.8** Fill in the blanks in the derivation of the tensor equation.

**(a)** Show that $\Gamma^i{}_{jk}$ is given by Eq. (6.57) in the presence of tensor perturbations.

**(b)** Show that the last term in Eq. (6.58) is given by Eq. (6.64).

**6.9** We defined the perturbation to the photon distribution function via Eq. (5.2). Show that, if $\Theta$ depends only on $\mu$, the cosine of the angle between $\hat{\boldsymbol{k}}$ and $\hat{\boldsymbol{p}}$, then $T^1{}_1 -$

$T^2{}_2$ vanishes if we choose $\hat{\boldsymbol{k}}$ to lie along the $z$-axis. This is indeed the dependence we have been dealing with so far. This is yet another aspect of the decomposition theorem: the terms that source the scalar perturbations (and are sourced by them) do not affect tensor perturbations.

**6.10** Show that scalar perturbations ($\Phi$ and $\Psi$) do not contribute to either $\delta G^1{}_1 - \delta G^2{}_2$ or to $\delta G^1{}_2$ if $\hat{\boldsymbol{k}}$ is along the $z$-axis. This completes the demonstration of the decomposition theorem for scalars and tensors.

**6.11** Use the ${}^1{}_2$ component of the Einstein equations to show that $h_\times$ obeys the same equation as does $h_+$.

**6.12** Solve the wave equation (6.73) if the universe is purely matter dominated. Do the same for the radiation-dominated case.

**6.13** Define the *transfer function* for gravitational-wave evolution as

$$T(k,\eta) \equiv \frac{h_t(k,\eta)}{h_t(k,\eta=0)}\left(\frac{k\eta}{3j_1(k\eta)}\right). \tag{6.84}$$

You might recognize the term in parentheses as the inverse of the matter-dominated solution you derived in Exercise 6.12. Solve Eq. (6.73) numerically in the fiducial cosmology and compute the transfer function at $\eta = \eta_0$.

**6.14** Derive the equation for the photon distribution function in the presence of tensor perturbations given by Eq. (6.49). Unlike scalar perturbations, tensor perturbations induce an azimuthal dependence in $\Theta_l$, so we need to decompose the anisotropy due to tensors into

$$\Theta^{\rm T}(k,\mu,\phi) = \Theta^{\rm T}_+(k,\mu)(1-\mu^2)\cos(2\phi) + \Theta^{\rm T}_\times(k,\mu)(1-\mu^2)\sin(2\phi). \tag{6.85}$$

Show that both the $+$ and the $\times$ component satisfy

$$\frac{d\Theta^{\rm T}_t}{d\eta} + ik\mu\Theta^{\rm T}_t + \frac{1}{2}h_t{}' = \tau'\left[\Theta^{\rm T}_t - \frac{1}{10}\Theta^{\rm T}_{t,0} - \frac{1}{7}\Theta^{\rm T}_{t,2} - \frac{3}{70}\Theta^{\rm T}_{t,4}\right] \tag{6.86}$$

where $t$ stands for either $\times$ or $+$, and the moments $\Theta^{\rm T}_{t,l}$ are defined in analogy to the scalar moments in Eq. (5.66).

**6.15** Consider vector perturbations to the metric. Specializing to $\hat{\boldsymbol{k}} = \hat{\boldsymbol{e}}_z$, these can be described by two function $h_{xz}$ and $h_{yz}$ where again only the spatial part of the metric is perturbed (recall from Sect. 6.2 that we can use the freedom of gauge choice to eliminate one of the two transverse vectors in the metric, in this case $B_i$). So, $h_{ij}$ is

$$h^V_{ij} = \begin{pmatrix} 0 & 0 & h_{xz} \\ 0 & 0 & h_{yz} \\ h_{xz} & h_{yz} & 0 \end{pmatrix}. \tag{6.87}$$

Relate $h_{xz}$, $h_{yz}$ to $V_i$, and show that $V_i$ is transverse. Then, show that $h_{xz}$ and $h_{yz}$ do not affect any of the equations we have derived so far for scalar or tensor evolution: Eq. (6.41), Eq. (6.48), and Eq. (6.73)—yet another aspect of the decomposition theorem.

# 7 Initial conditions

In previous chapters, with the goal of predicting the evolution of structure in the universe in mind, we have derived the equations governing perturbations around a smooth background. Before we start solving these equations though, we need to know the initial conditions. This quest for initial conditions leads to an entirely new realm of physics, the theory of inflation. Inflation was introduced (Guth, 1981; Sato, 1981; Linde, 1982; Starobinsky, 1982; Albrecht and Steinhardt, 1982) partly to explain how regions that could not have been in causal contact with each other (Fig. 7.1) have the same temperature—in other words, why the universe we live in is so homogeneous on large scales. It was soon realized that the very mechanism that explains the uniformity of the temperature can also account for the origin of perturbations in the universe. Therefore, understanding inflation will provide us with the initial conditions we need in order to solve the system of Einstein and Boltzmann equations. It is very difficult to test a theory like inflation directly, since the underlying physics might only show itself on energy scales well beyond the reach of accelerators. Nonetheless, it is by far the most plausible explanation for the seeds of structure, and will be put to increasingly stringent tests by the upcoming generation of CMB and large-scale structure surveys. One set of generic predictions of the inflation scenario has been verified experimentally: that the initial conditions for structure are Gaussian, adiabatic and nearly scale-invariant with a spectral index slightly less than one. We will learn what this means precisely in this chapter.

## 7.1 The horizon problem and a solution

The Einstein–Boltzmann system of equations we have derived is no different from most other problems in physics in that we need initial conditions in order to solve it—in our case, the initial conditions for the entire universe. Once we have these initial conditions, the future evolution is completely determined. Finally, then, it is time to confront a problem that we have politely ignored throughout this book so far. This problem, usually called the *horizon problem*, is: why is the universe so smooth and so big?

We have seen in Ch. 1 that, at an age of 380,000 years, the observable universe was very uniform, to roughly 1 part in $10^5$ (the typical amplitude of the temperature fluctuations in the CMB); moreover, the ingredients we can observe directly, photons and baryons, were extremely close to thermal. This clearly made our life much easier, for we could work in the limit of small perturbations around an FLRW universe. But why is this so? A generic patch of spacetime the size of the observable universe, i.e. a patch where we randomly choose the initial densities of matter and radiation at each point, would be highly inhomogeneous. The first possible explanation that comes to mind is thermalization: if we start

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00013-9
Copyright © 2021 Elsevier Inc. All rights reserved.

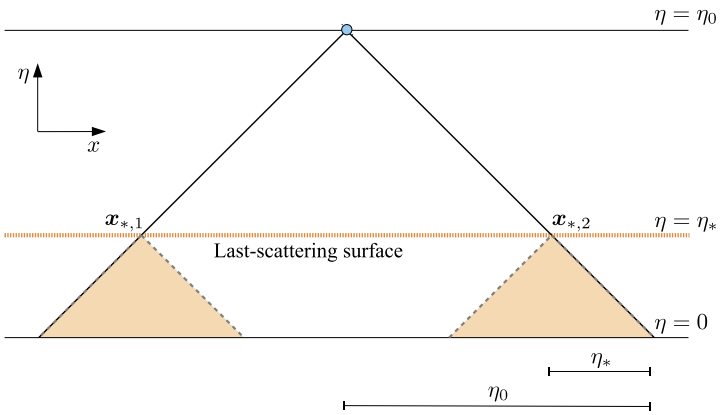



**FIGURE 7.1** The horizon problem illustrated in a diagram of $\eta$ vs. $x$, with the two other spatial dimensions $(y, z)$ suppressed. We as observers (top center) detect light signals coming from our past light cone (diagonal solid lines). The observed CMB is emitted when this cone intersects the last-scattering surface $\eta = \eta_*$ (horizontal dashed line), and is found to be uniform. Only signals from within the shaded regions below each point on the last-scattering surface could have influenced the CMB photons emitted from $\boldsymbol{x}_{*,1}$ and $\boldsymbol{x}_{*,2}$. Since these regions do not overlap, no form of causal physics could have allowed them to adjust to the same temperature if they started from different temperatures. This is because the comoving horizon $\eta_*$ at the time the CMB was emitted is much smaller than our comoving horizon now, $\eta_0$.

with a very inhomogeneous universe and let it be in thermal contact, then eventually the entire universe will equilibrate at the same, shared temperature, just as if we brought many containers with gas at various temperatures in contact with each other. In our universe, this solution does not seem to work, because different parts of the universe observed in the map of the CMB were so far apart at the time of recombination that they were not in causal contact with one another (Fig. 7.1): that is, even light could not have traveled from one region to another. Therefore, they were seemingly never able to thermalize.

We can quantify the extent of this problem by computing the *comoving horizon* $\eta_*$ at recombination: the comoving distance light could have traveled from $\eta = 0$ to $\eta_*$ (recall the discussion after Eq. (2.35)). We then compare that to the comoving distance separating two different patches that we see on the CMB sky today. In the concordance cosmological model, and assuming that the universe contained only matter and radiation all the way back to $t = 0$, we find that the comoving horizon at recombination is $\eta_* = \eta(a_*) \approx 281\,h^{-1}$ Mpc. The comoving distance between patches on the CMB sky today separated by an angle $\theta$ is (for small $\theta$)

$$\chi(\theta) \simeq \chi_* \theta = (\eta_0 - \eta_*)\theta. \tag{7.1}$$

Now, $\eta_0 \approx 14200\,h^{-1}$ Mpc, so that two patches in the CMB separated by

$$\theta \geq \frac{\eta_*}{\eta_0 - \eta_*} \approx 1.2^\circ \tag{7.2}$$

cannot have been in thermal contact at recombination. The large factor between $\eta_0$ and $\eta_*$ exacerbates the problem, which is therefore quite a bit more severe than the cartoon in Fig. 7.1 would suggest (where $\eta_0/\eta_*$ is of order 3, rather than the factor 50 of the real universe).

To gain insight into an assumption that underlies this problem, it is useful to rewrite Eq. (2.35) as an integral over the scale factor. Changing integration variables from $t'$ to $\ln a'$ leads to

$$\eta(a) = \int_0^a d\ln a' \frac{1}{a'H(a')}. \tag{7.3}$$

Thus, the comoving horizon $\eta$ is the logarithmic integral of the *comoving Hubble radius*, $1/aH$. The comoving Hubble radius is the approximate distance over which light can travel in the course of one expansion time, i.e., the time in which the scale factor increases by a factor of $e$. It provides a yardstick to assess whether particles can, at the given epoch, communicate within one $e$-fold of expansion. If the universe is dominated by either matter or radiation as we assumed, then $H$ scales as either $a^{-3/2}$ or $a^{-2}$ and the comoving Hubble radius is always increasing. Therefore the largest contribution to $\eta$ comes from the most recent epochs.

This points the way to a solution: if there was an early epoch during which the comoving Hubble radius *decreased*, then $\eta_*$ may have received large contributions from very early times when the Hubble radius was much larger. In such a case, the size of a region that is able to thermalize early on could have been much larger than we naively computed, and much larger than our current comoving horizon. Now, an epoch during which $(aH)^{-1} = \dot{a}^{-1}$ decreases corresponds to one of increasing $\dot{a}$, or $\ddot{a} > 0$: the condition for an accelerated expansion. So, **an epoch of early acceleration would solve the horizon problem**. This postulated epoch is called ***inflation***.

Consider then Fig. 7.2, which shows the comoving Hubble radius as a function of scale factor. The right half of this plot tells us that, going back in time, the comoving scales of interest to us were much larger than $1/aH$, and more so as you go back further in time. The left-hand side of the plot shows that an inflationary epoch reduces the comoving Hubble radius dramatically. At some early point during inflation, the comoving Hubble radius was very large, larger than any scale of cosmological interest today: all scales of interest were well within the horizon.

Fig. 7.3 gives a different view of how inflation affects the causality argument. Both panels show the same *physical* scale. The left panel shows the comoving grid at some point during inflation, with the circle indicating the size of the comoving Hubble radius at that time. All particles (depicted by dots) within that region are in causal contact with one another, and therefore that entire region could have thermalized. The right panel shows what happens after another factor $\sim 8$ of expansion (let us suppose this is at the end of inflation). The comoving grid has expanded, and now the comoving Hubble radius covers a factor of 8 fewer cells (in each dimension) on the comoving grid. It appears that only a small region on the grid is within $1/aH$, i.e. within causal contact now, but in fact we know from the left panel that a much larger region was in causal contact before, during earlier stages of

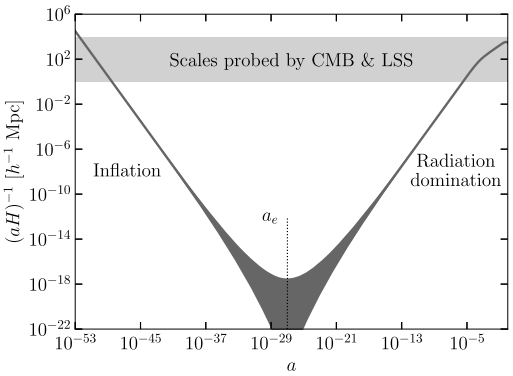



**FIGURE 7.2** The comoving Hubble radius as a function of scale factor. The main epochs are clearly visible: inflation at early times (with $H = H_{\rm inf}$ constant), transitioning to radiation domination around $a \sim a_e$, and finally matter domination at $a \gtrsim 10^{-4}$. Dark energy domination is barely visible as a further flattening at $a \gtrsim 0.5$. Scales of cosmological interest (horizontal shaded band) were larger than the Hubble radius when $a \lesssim 10^{-5}$. They later entered the Hubble radius where we are able to observe them. Very early on during inflation, however, all scales of interest were smaller than the Hubble radius and therefore within causal contact. The evolution of the Hubble radius around the end of inflation (at $a_e$) is uncertain, as indicated by the dark shaded band. However, since all modes of interest were far outside the horizon at that time, they are largely oblivious to the details of that epoch.

inflation. That is, after inflation, the patch of the universe that has been in causal contact is much larger than the comoving Hubble radius.

Yet another way to think about this is that inflation *empties out* the universe. As the universe expands exponentially, the particles in it get diluted accordingly. This is clear from Fig. 7.3: there are much fewer particles in a given physical volume in the right panel compared to the left. Let us assume that there is a substance that is keeping the Hubble rate $H = H_{\rm inf}$ approximately constant during inflation, a fact that is supported by the data, as we will see later. In that case, since $d\ln a = H\,dt$, the scale factor evolves as

$$a(t) = a_e e^{H_{\rm inf}(t-t_e)} \qquad (t < t_e) \tag{7.4}$$

where $t_e$ is the time at the end of inflation. As inflation proceeds, the universe becomes dominated by the smooth substance that is driving the acceleration, turning a chaotic, inhomogeneous patch of the universe into a *much* larger space that is completely smooth and empty. Imagine such a patch of the universe near the beginning of inflation, as depicted in the bottom left of Fig. 7.4. Once inflation sets in, anything that was in this patch—heavy particles, massless particles, magnetic monopoles, and so forth—soon becomes irrelevant, because it is rapidly diluted; the number density of massive particles evolves as $n(t) \propto a^{-3} \propto \exp(-3H_{\rm inf}t)$, so it drops exponentially fast, while the energy density driving inflation remains approximately constant. This is illustrated in the upper panel in Fig. 7.4. Along with the components of ordinary matter and radiation, perturbations to the spacetime are also quickly smoothed out. This is not unlike what happens when inflating a balloon: initially wrinkly, its surface becomes increasingly smooth as it is inflated (see also Exercise 7.1).

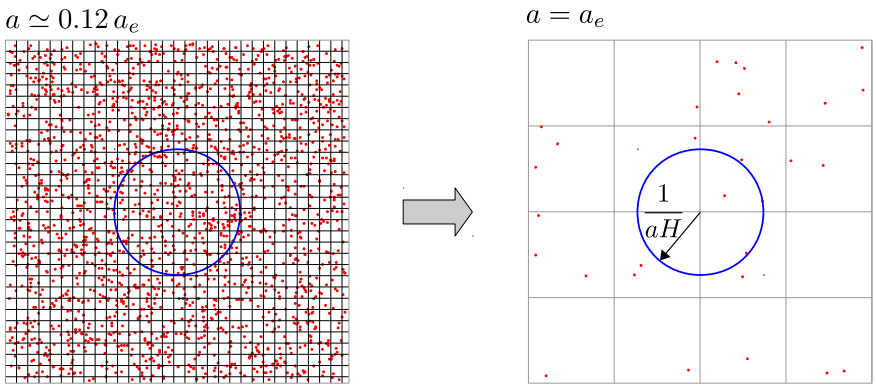



**FIGURE 7.3** Particles on the comoving grid during (left) and at the end of inflation (right). Both panels use the same *physical* (not comoving) scale. The dots illustrate particle positions, while the circles show the size of the comoving Hubble radius. At some time during inflation, the comoving Hubble radius was quite large, encompassing dozens of cells on the grid. After inflation, the comoving Hubble radius has shrunk to just one cell. Notice the much smaller number of particles within the comoving Hubble radius on the right. In this caricature, the scale factor has grown by a factor of order 8; during inflation the scale factor increases by a factor of order $e^{60}$. The shrinking of the comoving Hubble radius means that particles which were initially in causal contact with one another (within the circle on the left) can now no longer communicate within an $e$-fold of expansion. The physical Hubble radius $H^{-1}$ on the other hand remains roughly constant during inflation.

The universe at the end of inflation (right-most panel in Fig. 7.4) is smooth but completely empty. Fortunately, physical scenarios for inflation come with a built-in way to produce the desired hot Big Bang universe: the substance that drives the exponential expansion dominates the energy everywhere, and it is virtually the same everywhere. This energy is converted to ordinary particles, which quickly thermalize. Since the energy density was the same everywhere in the universe, the temperature is likewise uniform.

We used the qualifier "virtually" above because small perturbations are in fact generated during inflation. A key epoch in the evolution of a perturbation of comoving wavenumber $k$ is when its comoving wavelength becomes of order the comoving Hubble radius $(aH)^{-1}$. During radiation and matter domination (indeed, any time when $1/aH$ is growing), the modes always evolve from $k^{-1} \gg (aH)^{-1}$ initially to $k^{-1} < (aH)^{-1}$ at later times (see the right half of Fig. 7.2, where the relevant wavelengths are depicted by the horizontal band). We say that the mode *enters the horizon* as it goes from $k \ll aH$ to $k \gtrsim aH$, since it becomes an observable perturbation for an observer living in the universe. The equations we developed in the previous chapters allow us to follow the perturbations from outside the horizon until the time we observe them.

Inflation adds a mirror image of this behavior at early times (left side of Fig. 7.2): modes initially have $k \gg aH$, but then *leave the horizon* since $aH$ shrinks exponentially, so that $k \ll aH$ at the end of inflation for all modes that we can possibly observe directly. Fig. 7.2 shows that the largest observable scales today, those which entered the horizon very recently, left the horizon earliest. Small scales which entered the horizon a long time ago exited correspondingly later during inflation. To explain the structure in the universe to-

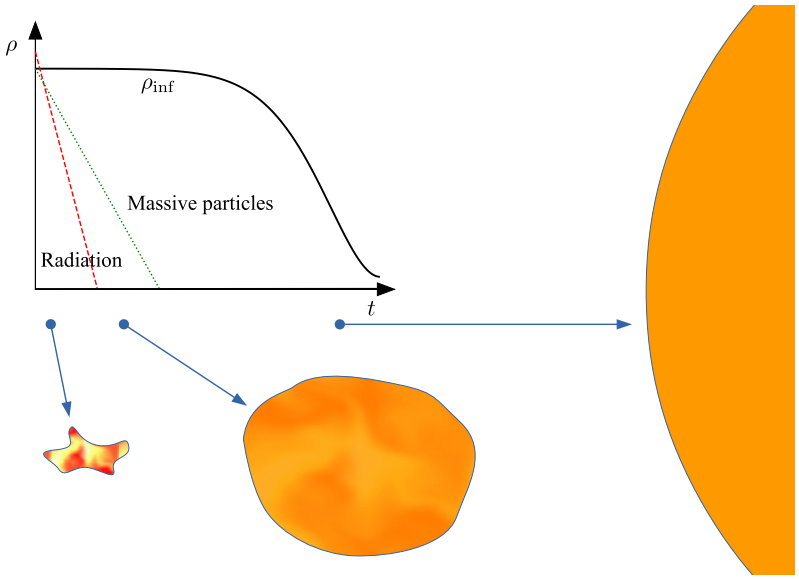



**FIGURE 7.4** The universe empties out and becomes homogeneous as inflation progresses. The upper left plot schematically shows the evolution of the energy density of the different components (in arbitrary logarithmic units). As inflation proceeds, any primordial radiation and massive particles are quickly diluted, and only the substance driving inflation remains (the black line shows its energy density $\rho_{\rm inf}$, which is constant during inflation). At the end of inflation, the substance decays into radiation and massive particles again (not shown). The sketches at the bottom and right illustrate the patch corresponding to the observable universe at different stages of inflation (in physical units, but not to scale). The spacetime patch started out in a highly inhomogeneous and curved state, and became more homogeneous and very close to Euclidean as the exponential expansion proceeded.

day, then, it is clearly important to understand the generation of perturbations during inflation.

Before we turn to actual physical models of inflation, let us understand how much accelerated expansion is required to solve the horizon problem. How long should inflation last? Solving the horizon problem requires the comoving Hubble radius to have been larger than the current comoving radius, $H_0^{-1}$, before inflation. The comoving Hubble radius at the end of inflation, at time $t_e$, was $1/a_e H_e$ where $H_e \equiv H(t_e)$. To get a rough order-of-magnitude estimate, we will assume that the temperature after inflation was $T_e = 10^{14}$ GeV and ignore the relatively brief epochs of matter and recent dark energy domination, so that we can use the relations of radiation domination throughout (you can correct this in Exercise 7.3). Then, $H$ scales as $a^{-2}$, and the ratio of the comoving Hubble radius at the end of inflation and today is $a_0 H_0/a_e H_e = a_e/a_0$. Since $T \propto a^{-1}$, $a_e/a_0 \simeq T_0/10^{14}$ GeV $\simeq 10^{-27}$. So the comoving Hubble radius at the end of inflation was 27 orders of magnitude smaller than it is today. We conclude that the scale factor had to increase by a factor of $10^{27} \simeq e^{62}$ during inflation in order for the current comoving Hubble radius to be smaller than that at the beginning of inflation. If the Hubble rate is constant, then the expansion is exponen-

tial in time, and we say that the universe needed to expand exponentially for roughly 60 $e$-folds.

One final technical point: the total comoving horizon $\eta$ as given in Eq. (7.3) ceases to be an effective time parameter after inflation because it becomes very large very early on and then changes relatively little as the universe expands during the matter- and radiation-dominated eras. A common way to deal with this issue is to redefine $\eta$ such that $\eta = 0$ corresponds to the end of inflation:

$$\eta(t) = \int_{t_e}^{t} \frac{dt'}{a(t')}, \tag{7.5}$$

where $t$ could be larger or smaller than $t_e$. Note that this means that during inflation, $\eta$ is negative, but always monotonically increasing. This definition is convenient, because we never need to refer to the starting time of inflation (it could last for much more than 60 $e$-folds); the end time $\eta = 0$ is all that matters.

To sum up, inflation—an epoch in which the universe expands exponentially—solves the horizon problem. At the end, the homogeneous substance driving inflation converts into other particles, producing a homogeneous hot universe filled with matter and radiation. In addition, inflation is guaranteed to produce some amount of fluctuations around the homogeneous universe, the minimal amount of fluctuations guaranteed by Heisenberg's uncertainty principle. One of the main goals of this chapter is to describe the generation and evolution of perturbations during inflation.

## 7.2 Inflation

We already know from our study of dark energy in Sect. 2.4.6 that in order to obtain an accelerating expansion, the effective pressure must be negative. Inflation was apparently driven by a similar form of energy, one with $\mathcal{P} < 0$. To reiterate what we emphasized in Ch. 2, negative pressure is not something with which we have any familiarity. Nonrelativistic matter has small positive pressure, while a relativistic gas has $\mathcal{P} = \rho/3$, again positive. So whatever it is that drives inflation (or the recent acceleration of the universe) is not ordinary matter or radiation. It cannot be a cosmological constant either: a cosmological constant would lead to perpetual rapid inflation, while we need inflation to end and transition to the radiation- and then matter-dominated phases which we observe.

The simplest possibility to generate such a transitory epoch of accelerated expansion is via the potential energy of a scalar field (incidentally, this is also precisely what underlies a class of models of dark energy dubbed *quintessence*). It is worth noting that there is no known scalar field that can drive inflation. While we know of a scalar field in nature, the Higgs boson (Box 1.1), using it to drive inflation turns out to be difficult. Its interactions and properties are sufficiently constrained by now for us to know that it cannot serve as the source for inflation. On the other hand, most proposed fundamental particle physics theories, such as string theory implementations, contain additional scalar fields. Indeed, the vast majority of current work on inflation is based on a scalar field (or sometimes several).

Hence, we will do the same and drop any pretensions of connecting the generic scalar field we employ to drive inflation to known physics.[1] Keep in mind, however, that it may well be true that the idea of inflation is correct but it is driven by something other than a scalar field.

An even more radical alternative, named ekpyrosis, involves a slowly contracting phase of the universe instead of the rapid acceleration posited by inflation. In order to transition to the presently observed expanding universe, one has to go through a "bounce" where the Hubble rate changes sign. This bounce turns out to be difficult to control. We will not describe the ekpyrotic scenario here, but point out that the techniques used for calculating the perturbations in ekpyrosis are quite similar to those used for inflationary calculations.

We want to know if a scalar field—which we will call $\phi(\boldsymbol{x}, t)$, not to be confused with the metric perturbation $\Phi(\boldsymbol{x}, t)$—can have negative $\rho + 3\mathcal{P}$. So our first task is to write down the energy-momentum tensor for $\phi$. This can be derived from the Lagrangian for a canonical scalar field with a potential; see Exercise 7.4. It is

$$T^{\alpha}{}_{\beta} = g^{\alpha\nu}\frac{\partial\phi}{\partial x^{\nu}}\frac{\partial\phi}{\partial x^{\beta}} - \delta^{\alpha}{}_{\beta}\left[\frac{1}{2}g^{\mu\nu}\frac{\partial\phi}{\partial x^{\mu}}\frac{\partial\phi}{\partial x^{\nu}} + V(\phi)\right]. \tag{7.6}$$

Here $V(\phi)$ is the potential for the field. For example, a free field with mass $m$ has a potential $V(\phi) = m^2\phi^2/2$. A warning about signs: if you delve into the literature you will invariably find different signs than those in Eq. (7.6). These are dictated by the choice of metric. Although our metric signature $(-,+,+,+)$ is the most common convention in the context of cosmology, it is not as common in particle physics where the "mostly negative" convention is more popular. We will assume that the field is homogeneous to zeroth order, consisting of a zeroth-order part and a first-order perturbation, $\delta\phi(\boldsymbol{x}, t)$. In this section we will derive information about the zeroth-order homogeneous part, $\phi(t)$, specifically its energy density and pressure as well as its time evolution. Later we will consider the perturbations $\delta\phi$, and how they are generated.

For the homogeneous part of the field, only time derivatives of $\phi$ are relevant, so the indices $\nu$, $\beta$ in the first term in Eq. (7.6) and $\mu$, $\nu$ in the second must be equal to zero. The energy-momentum tensor then reduces to

$$T^{\alpha}{}_{\beta} = -\delta^{\alpha}{}_{0}\delta^{0}{}_{\beta}\dot{\phi}^2 + \delta^{\alpha}{}_{\beta}\left[\frac{1}{2}\dot{\phi}^2 - V(\phi)\right]. \tag{7.7}$$

The time-time component of $T^{0}{}_{0}$ is equal to $-\rho$, so the energy density is

$$\rho = \frac{1}{2}\dot{\phi}^2 + V(\phi). \tag{7.8}$$

The first term here is the kinetic energy density of the field, the second its potential energy density. A homogeneous scalar field therefore has much the same dynamics as a single

[1] Making this connection is left as a homework problem for a future Nobel laureate.

particle moving in a potential: think of $\phi(t)$ as the position of the particle $x(t)$, and $\dot{\phi}$ as its velocity $\dot{x}$. In fact this analogy even enters the language used to describe inflation. The pressure for the homogeneous field is $\mathcal{P} = T^i{}_i$ (no sum over spatial index $i$; it is the same for $i = 1, 2, 3$), so

$$\mathcal{P} = \frac{1}{2}\dot{\phi}^2 - V(\phi). \tag{7.9}$$

A field configuration with negative pressure is therefore one with more potential energy than kinetic. This is equivalently phrased as an equation of state

$$w = \frac{\mathcal{P}}{\rho} = \frac{\dot{\phi}^2 - V(\phi)}{\dot{\phi}^2 + V(\phi)} \tag{7.10}$$

that is close to $-1$.

The most popular scenario of inflation assumes a scalar field slowly rolling toward its true ground state(Linde, 1982; Albrecht and Steinhardt, 1982). The potential energy of such a field is very close to constant (if the potential is not too steep) so it quickly comes to dominate over the kinetic energy (and the energy of all other particles). An example is shown in Fig. 7.5; inflation ends once the field has reached the minimum of the potential, where it will oscillate and decay into lighter particles (and eventually those of the standard model). Many different forms of potentials have been proposed in the literature; see Martin et al. (2014) for an exhaustive list. Fortunately or not, many different models can be made to fit the observations. For this reason, we will not discuss individual scenarios in detail here.

To determine the evolution of $\phi$ for any potential, consider the conservation of the energy-momentum tensor:

$$\nabla_\mu T^\mu{}_\nu = \frac{\partial T^\mu{}_\nu}{\partial x^\mu} + \Gamma^\mu{}_{\alpha\mu} T^\alpha{}_\nu - \Gamma^\alpha{}_{\nu\mu} T^\mu{}_\alpha = 0. \tag{7.11}$$

The stress-energy tensor for the homogeneous background field $\phi(t)$ is of the same form as Eq. (2.44), so that we can use our result from Ch. 2, Eq. (2.56):

$$\frac{\partial \rho}{\partial t} + 3H[\rho + \mathcal{P}] = 0. \tag{7.12}$$

Applying this to the density and pressure we obtained above yields

$$\dot{\phi}\ddot{\phi} + V_{,\phi}\,\dot{\phi} + 3H\dot{\phi}^2 = 0, \tag{7.13}$$

where here and throughout $V_{,\phi} \equiv dV/d\phi$, and, upon dividing by $\dot{\phi}$,

$$\ddot{\phi} + 3H\dot{\phi} + V_{,\phi}(\phi) = 0. \tag{7.14}$$

Let us switch to conformal time $\eta$ as time variable; then it is straightforward to show that (Exercise 7.5)

$$\phi'' + 2aH\phi' + a^2 V_{,\phi} = 0. \tag{7.15}$$

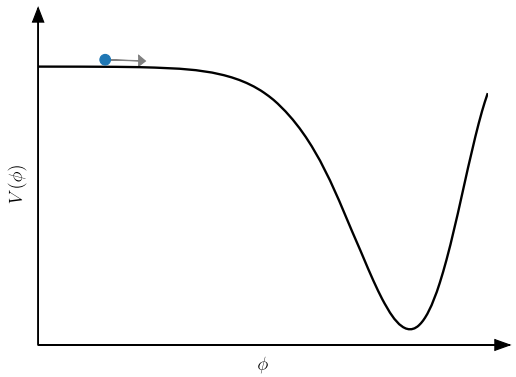



**FIGURE 7.5** A scalar field slowly rolling down a potential $V(\phi)$. Since it rolls slowly, it has little kinetic energy. The potential energy is nonzero, however, so the pressure is negative. The inflationary epoch ends once the field has reached the minimum of the potential.

It is worth emphasizing that quite literally all results so far directly apply to dark energy ("quintessence") models described by a canonical scalar field as well.

Most models of inflation are *slow-roll* models, in which the zeroth-order field, and hence the Hubble rate, vary slowly. Therefore, a simple relation between the conformal time $\eta$ and the expansion rate holds. In particular, during inflation

$$\eta \equiv \int_{a_e}^{a} \frac{da}{Ha^2} \simeq \frac{1}{H}\int_{a_e}^{a} \frac{da}{a^2} \simeq -\frac{1}{aH} \tag{7.16}$$

where the first approximate equality holds because $H$ is nearly constant, and the second because the scale factor at the end of inflation is much larger than in the middle ($a_e \gg a$). To quantify slow roll, cosmologists typically define two variables that vanish in the limit that $\phi$ remains constant. Several different options and conventions exist. We will focus on the one most directly linked to observables; you can derive the relation to other choices in Exercise 7.7 and Exercise 7.8. First,

$$\epsilon_{\rm sr} \equiv \frac{d}{dt}\left(\frac{1}{H}\right) = -\frac{H'}{aH^2}, \tag{7.17}$$

which yields the fractional change in the Hubble rate during one $e$-fold of expansion. Since $H$ is always decreasing, $\epsilon_{\rm sr}$ is always positive. During inflation, it is typically small, whereas it is equal to 2 during the radiation era. In fact, one definition of an inflationary epoch is one in which $\epsilon_{\rm sr} < 1$; technically speaking, it quantifies the departure of the spacetime from an exact de Sitter space, an empty universe with a positive cosmological constant.

The second variable which quantifies how slowly the field is rolling is

$$\delta_{\rm sr} \equiv \frac{1}{H}\frac{\ddot{\phi}}{\dot{\phi}} = -\frac{1}{aH\phi'}\left[aH\phi' - \phi''\right]$$
$$= -\frac{1}{aH\phi'}\left[3aH\phi' + a^2 V_{,\phi}\right]. \tag{7.18}$$

The second line follows from Eq. (7.15). In the literature, the slow-roll parameters are usually denoted simply as $\epsilon$, $\delta$, but due to the paucity of Greek letters we opt to make them unambiguous with the "sr" subscript. Again, in most inflationary models $\delta_{\rm sr}$ is small. We will see later why the landscape of slow-roll inflation models is described precisely by two parameters; in particular, they quantify the key features of inflationary predictions, namely deviations from the simplest possible spectrum of perturbations and the production of gravitational waves.

The slow-roll phase cannot last indefinitely, since inflation must end at some point. This point is reached when the potential steepens and the field reaches the potential minimum depicted in Fig. 7.5. At that point, the field is no longer slowly rolling, but has significant kinetic energy, so it oscillates around the minimum. Then, the equation of state Eq. (7.10) is no longer close to $-1$, but close to zero, so that the universe has transitioned to an epoch of decelerated expansion. Then, finally, $\phi$ decays into lighter particles. Eventually, perhaps after a long chain of decays, the result is an almost completely homogeneous, radiation-dominated universe. The highly uncertain details of this transition from inflation to decelerated expansion, known as *reheating*, do not affect the perturbations we are interested in, since they are far outside the horizon at the end of inflation. They remain frozen throughout this transition.

## 7.3 Gravitational wave production

Inflation does more than solve the horizon problem. The power of inflation is its ability to correlate scales that would otherwise be disconnected. The zeroth-order scheme outlined in the previous section ensures that the universe will be close to uniform on all scales of interest today. There are perturbations about this zeroth-order scheme, though, and these perturbations—produced early on when the scales are causally connected—persist long after inflation has terminated.

We are most interested in *scalar* perturbations to the metric since these couple to the density of matter and radiation and ultimately are responsible for the structure we observe in the universe. In Sect. 7.4 we will study these in detail. In addition to scalar perturbations, though, inflation also generates *tensor* fluctuations in the metric, that is, gravitational waves. As we saw in Ch. 6, these are not coupled to the density and so are not responsible for the large-scale structure of the universe, but they do induce anisotropies in the CMB. In fact, these anisotropies turn out to be a unique signature of inflation and offer the best window on the physics driving inflation, so they are clearly worthy of our study. We choose to study the production of tensor perturbations first, because they are simpler than

scalar perturbations: in Sect. 6.2 we saw that tensor perturbations are gauge invariant; that is, they look the same regardless of which coordinates we choose. Moreover, we can neglect the coupling of tensor modes to the other perturbations in the metric or matter. Scalar perturbations are afflicted by both of these complications: they look different depending on which coordinate system is chosen, and the perturbations of the scalar field mix with the perturbations in the metric. So in order not to obfuscate the main point of the generation of perturbations out of vacuum fluctuations during inflation, we study tensor modes first.

During inflation, the universe consists primarily of a uniform scalar field and a uniform background metric. Against this background, the fields fluctuate quantum-mechanically. At any given time, the average fluctuation is zero, because there are regions in which the field is slightly larger than its average value and regions in which it is smaller. The average of the square of the fluctuations (the variance), however, is not zero. Our goal is to compute this variance and see how it evolves as inflation progresses. Looking ahead, once we know this variance, we can draw from a distribution with this variance to set the initial conditions with which to start the calculation of the evolution of structure (Sect. 7.5).

In cosmology, we always work in terms of statistics, such as the correlation function and power spectrum, because no known theory predicts the overdensity in a given spot on the sky. In the inflationary scenario, this uncertainty is fundamental: inflation erases all traces of what came before it, and replaces those with quantum-mechanical vacuum fluctuations, which cannot be predicted *in principle*. What inflation predicts then is precisely the statistical distributions from which the perturbations are drawn.

This chapter is the first, and only, chapter in which we will encounter quantum field theory. Field theory has a reputation as a difficult subject which is not entirely undeserved, but the part we will need for inflation is closely connected to ordinary quantum mechanics and therefore relatively straightforward. As a warmup, we will consider the quantization of the one-dimensional harmonic oscillator.

### 7.3.1 Quantizing the harmonic oscillator

In order to compute the quantum fluctuations in the metric, we need to quantize the field. The way to do this, in the case of both tensor and scalar perturbations, is to rewrite the problem so that it looks like a simple harmonic oscillator. Once that is done, we will appeal to our knowledge of this simple system. Therefore, let us first record some basic facts about the quantization of the harmonic oscillator.[2]

- A harmonic oscillator with frequency $\omega$ is governed by the equation

$$\frac{d^2x}{dt^2} + \omega^2 x = 0. \tag{7.19}$$

[2] To those recalling their quantum mechanics classes: we will be working in the Heisenberg picture throughout, where states are fixed but operators evolve.

- Upon quantization, $x$ becomes a quantum operator

$$\hat{x} = v(\omega, t)\hat{a} + v^*(\omega, t)\hat{a}^\dagger \tag{7.20}$$

  where $\hat{a}$ is the *annihilation* operator, and $v$ is the positive-frequency solution to Eq. (7.19), $v \propto e^{-i\omega t}$. A dagger on operators denotes their Hermitian conjugate.
- $\hat{a}$ annihilates the *vacuum* state $|0\rangle$ (hence its name), in which there are no particles. It also satisfies the commutation relation

$$[\hat{a}, \hat{a}^\dagger] \equiv \hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = 1. \tag{7.21}$$

  Other commutators vanish: $[\hat{a}, \hat{a}] = [\hat{a}^\dagger, \hat{a}^\dagger] = 0$. Since $a^\dagger$ acting on the vacuum state leads to a state containing a single particle, it is called the *creation* operator. It is straightforward to show (Exercise 7.10) that these commutation relations are equivalent to the following commutation relations between the position and momentum operators $\hat{x}$ and $\hat{p}$:

$$[\hat{x}, \hat{p}] = i, \tag{7.22}$$

  as long as $v$ is normalized via

$$v(\omega, t) = \frac{e^{-i\omega t}}{\sqrt{2\omega}}. \tag{7.23}$$

These facts enable us to compute the quantum fluctuations of the operator $\hat{x}$ in the ground state $|0\rangle$:

$$\begin{aligned}\left\langle |\hat{x}|^2 \right\rangle &\equiv \langle 0|\hat{x}^\dagger\hat{x}|0\rangle \\ &= \langle 0|(v^*\hat{a}^\dagger + v\hat{a})(v\hat{a} + v^*\hat{a}^\dagger)|0\rangle.\end{aligned} \tag{7.24}$$

We will later identify $\hat{x}$ with the field $\phi$, which we have seen behaves like the position of a particle in a potential well. Notice that here we use the notation $\langle\hat{X}\rangle$ to denote the *vacuum expectation value* of the operator $\hat{X}$. In later chapters we will use the same notation to denote *ensemble averages* of observables, that is, the value obtained for an observable if it were to be measured over an infinite volume. These two are not the same. However, we can equate them if our universe is only a small part of a larger patch of spacetime that underwent inflation. This subtle identification underlies all of modern cosmology.

Since $\hat{a}|0\rangle = 0$, the first term in the second set of parentheses vanishes. Similarly, $\langle 0|\hat{a}^\dagger = (a|0\rangle)^\dagger = 0$, so we are left with

$$\begin{aligned}\left\langle |\hat{x}|^2 \right\rangle &= |v(\omega,\ t)|^2\langle 0|\hat{a}\hat{a}^\dagger|0\rangle \\ &= |v(\omega, t)|^2\langle 0|[\hat{a}, \hat{a}^\dagger] + \hat{a}^\dagger\hat{a}|0\rangle.\end{aligned} \tag{7.25}$$

The second term again vanishes since $\hat{a}$ annihilates the vacuum, while the first is unity, so the variance in $\hat{x}$ is

$$\left\langle |\hat{x}|^2 \right\rangle = |v(\omega,\ t)|^2, \tag{7.26}$$

which evaluates to $1/(2\omega)$. This is (almost) all we need to know about quantum fluctuations in order to compute the fluctuations in the early universe generated by inflation.

Before delving into the calculation, let us briefly give an intuitive picture of the generation of perturbations during inflation. Instead of dealing with a single harmonic oscillator, we will now deal with an infinite collection of oscillators, one for every Fourier mode $\boldsymbol{k}$. Each mode carries its own individual creation and annihilation operators $\hat{a}^\dagger_{\boldsymbol{k}}$, $\hat{a}_{\boldsymbol{k}}$. The time evolution of these operators is described by a combination of positive and negative frequencies, which in Minkowski space simply are $v(\boldsymbol{k},t)$, $v^*(\boldsymbol{k},t) \propto \exp(\pm i\omega(\boldsymbol{k})t)$. In Minkowski space, then, the vacuum expectation value Eq. (7.26) is independent of time and position, and can be subtracted: no real particles are produced.

This changes in a rapidly expanding spacetime. As we will see, the two independent solutions out of which we assemble the mode functions $v(\boldsymbol{k},\eta)$ have drastically different time dependences. Physically, the vacuum state is evolving due to the expansion, so that the vacuum state at the beginning of inflation is no longer devoid of particles later on. In the case we will study first, the particles produced are the gravitons that form gravitational waves. The variance of the fluctuations will be identified as the power spectrum of gravitational waves.

### 7.3.2 Tensor perturbations

Recall that tensor perturbations to the metric are described by two functions $h_+$ and $h_\times$, each of which obeys Eq. (6.73),

$$h'' + 2\frac{a'}{a}h' + k^2 h = 0 \qquad (h = h_+,\ h_\times). \tag{7.27}$$

We consider a single tensor-mode polarization $t = +, \times$ in the following, but drop the subscript $t$ for clarity.

We would like to massage this equation into the form of a harmonic oscillator, so that $h$ can be easily quantized. To do this, define

$$\mathfrak{h} \equiv \frac{ah}{\sqrt{16\pi G}}. \tag{7.28}$$

We will see that the factor of $a$ leads to an equation for $\mathfrak{h}$ akin to the harmonic oscillator. Why the factor of $1/\sqrt{16\pi G}$? In order to obtain this factor, one has to derive the action for tensor perturbations in Minkowski space, which necessitates going to second order in perturbations without equipping us with any essential physics we will later need. Hence we take Eq. (7.28) as a given.

Derivatives of $h$ with respect to conformal time can be rewritten as

$$\frac{h'}{\sqrt{16\pi G}} = \frac{\mathfrak{h}'}{a} - \frac{a'}{a^2}\mathfrak{h} \tag{7.29}$$

and

$$\frac{h''}{\sqrt{16\pi G}} = \frac{\mathfrak{h}''}{a} - 2\frac{a'}{a^2}\mathfrak{h}' - \frac{a''}{a^2}\mathfrak{h} + 2\frac{(a')^2}{a^3}\mathfrak{h}. \tag{7.30}$$

Inserting these into Eq. (7.27), and getting rid of the factor $\sqrt{16\pi G}$, yields

$$\begin{aligned}\frac{\mathfrak{h}''}{a} - 2\frac{a'}{a^2}\mathfrak{h}' - \frac{a''}{a^2}\mathfrak{h} + 2\frac{(a')^2}{a^3}\mathfrak{h} + 2\frac{a'}{a}\left(\frac{\mathfrak{h}'}{a} - \frac{a'}{a^2}\mathfrak{h}\right) + k^2\frac{\mathfrak{h}}{a} \\ = \frac{1}{a}\left[\mathfrak{h}'' + \left(k^2 - \frac{a''}{a}\right)\mathfrak{h}\right] = 0.\end{aligned} \tag{7.31}$$

This is precisely the form we know how to use. It only involves $\mathfrak{h}$ and $\mathfrak{h}''$, analogous to Eq. (7.19), so we can immediately write down an expression for the quantum operator

$$\hat{\mathfrak{h}}(\boldsymbol{k}, \eta) = v(k, \eta)\hat{a}_{\boldsymbol{k}} + v^*(k, \eta)\hat{a}^\dagger_{\boldsymbol{k}}, \tag{7.32}$$

where the coefficients of the creation and annihilation operators satisfy

$$v'' + \left(k^2 - \frac{a''}{a}\right)v = 0. \tag{7.33}$$

We will shortly solve Eq. (7.33), but first let us see how the eventual solution determines the power spectrum of the fluctuations of the tensor perturbations. Using our harmonic oscillator analogy, Eq. (7.26), we can write the variance of perturbations in the $\mathfrak{h}$ field as

$$\langle\hat{\mathfrak{h}}^\dagger(\boldsymbol{k}, \eta)\hat{\mathfrak{h}}(\boldsymbol{k}', \eta)\rangle = |v(\boldsymbol{k}, \eta)|^2(2\pi)^3\delta^{(3)}_{\rm D}(\boldsymbol{k} - \boldsymbol{k}'). \tag{7.34}$$

Again, this is a vacuum expectation value of a quantum operator, which we will later identify with the ensemble average of a classical field. As we mentioned above, a quantum field is defined in all space, so it can be considered as an infinite collection of oscillators, each at a different spatial position, or, in Fourier space, at different values of $\boldsymbol{k}$. The quantum fluctuations in each of these oscillators are independent (as long as the equations are linear) so $\hat{\mathfrak{h}}(\boldsymbol{k})$ is completely uncorrelated with $\hat{\mathfrak{h}}(\boldsymbol{k}')$ if $\boldsymbol{k} \neq \boldsymbol{k}'$. The Dirac delta function in Eq. (7.34) enforces this independence; the $(2\pi)^3$ accounts for the fact that we work in the continuum limit. Recalling that $\mathfrak{h} = ah/\sqrt{16\pi G}$, we see that

$$\begin{aligned}\langle\hat{h}^\dagger(\boldsymbol{k}, \eta)\hat{h}(\boldsymbol{k}', \eta)\rangle &= \frac{16\pi G}{a^2}|v(k, \eta)|^2(2\pi)^3\delta^{(3)}_{\rm D}(\boldsymbol{k} - \boldsymbol{k}') \\ &\equiv P_h(k, \eta)(2\pi)^3\delta^{(3)}_{\rm D}(\boldsymbol{k} - \boldsymbol{k}')\end{aligned} \tag{7.35}$$

where the second line defines the power spectrum of the primordial tensor perturbations (for a single polarization). A related useful quantity is the dimensionless power spectrum

$$\Delta^2_h(k, \eta) \equiv \frac{k^3}{2\pi^2}P_h(k, \eta), \tag{7.36}$$

which gives the variance of tensor modes in a logarithmic wavenumber interval. With our definition,

$$P_h(k,\eta) = 16\pi G \frac{|v(k,\eta)|^2}{a^2}. \tag{7.37}$$

We have now reduced the problem of determining the spectrum of tensor perturbations produced during inflation to one of solving a second-order differential equation for $v(k,\eta)$, Eq. (7.33). To solve this equation, we first need to evaluate $a''/a$ during inflation. Recall that primes denote derivatives with respect to conformal time, so $a' = a^2 H \simeq -a/\eta$ by virtue of the last approximate equality in Eq. (7.16). Therefore, the second derivative of $a$ in Eq. (7.33) is

$$\begin{aligned} \frac{a''}{a} &\simeq -\frac{1}{a}\frac{d}{d\eta}\left(\frac{a}{\eta}\right) \\ &\simeq \frac{2}{\eta^2}. \end{aligned} \tag{7.38}$$

So the equation for $v$ is

$$v'' + \left(k^2 - \frac{2}{\eta^2}\right) v = 0. \tag{7.39}$$

The initial conditions necessary to solve this equation come from considering $v$ at very early times; specifically, when $k|\eta| \gg 1$, and the mode is "far inside the horizon." At that point, the $k^2$ term dominates, and the equation reduces precisely to that of the simple harmonic oscillator. In that case, we know (Eq. (7.23)) that the properly normalized solution is $e^{-ik\eta}/\sqrt{2k}$. This knowledge enables us to choose the proper solution to Eq. (7.39) (Exercise 7.12),

$$v = \frac{e^{-ik\eta}}{\sqrt{2k}}\left[1 - \frac{i}{k\eta}\right]. \tag{7.40}$$

This obviously goes into the correct solution when the mode is well within the horizon ($k|\eta| \gg 1$). Even if you do not work through Exercise 7.12, you should at least check that Eq. (7.40) is indeed a solution to Eq. (7.39).

After inflation has worked for sufficiently many $e$-folds, $k|\eta|$ becomes very small: the mode has exited the horizon. Taking the small-argument limit of Eq. (7.40), we have

$$\lim_{-k\eta\to 0} v(k,\eta) = \frac{e^{-ik\eta}}{\sqrt{2k}}\frac{-i}{k\eta}. \tag{7.41}$$

Recall that this corresponds to the evolution of $\mathfrak{h} \propto ah$. Hence, at early times $h \propto v/a$ falls as $1/a$ as inflation reduces the amplitude of the modes. Once $-k\eta$ becomes smaller than unity, the mode leaves the horizon, after which $h$ remains constant (since $1/\eta \propto a$), and becomes an observable gravitational wave once it re-enters the horizon. This production of gravitational waves is a consequence of the fact that the two solutions of the wave equation split into a constant and a decaying mode in an exponentially expanding spacetime.

Since we have normalized $v$, we can now determine the variance of the super-horizon gravitational-wave amplitude, which scales as $|v|^2/a^2$. It is constant in time after inflation has stretched the mode to be larger than the horizon. This constant determines the initial conditions for the gravitational waves, those with which to start off $h_+$, $h_\times$ at some time before the mode re-enters the horizon. Equations (7.37) and (7.41) show that this constant is

$$\begin{aligned} P_h(k) &= \frac{16\pi G}{a^2}\frac{1}{2k^3\eta^2} \\ &= \frac{8\pi G H^2}{k^3}\Big|_{k|\eta|=1}. \end{aligned} \tag{7.42}$$

The second line here follows from Eq. (7.16). This is our final expression for the primordial power spectrum of gravitational waves. We have assumed that $H$ is constant in deriving this result; in reality, $H$ varies slowly during inflation, but the result remains accurate if $H$ is evaluated at the time when the mode of interest leaves the horizon, $k|\eta| = 1$. Further, as noted at the beginning of this section, Eq. (7.42) is the power spectrum for $h_+$ and $h_\times$ individually; these are uncorrelated, so the total tensor power spectrum is larger by a factor of 2 (we will come back to this in Sect. 7.6).

A detection of these waves and measurement of $P_h(k)$ then would, quite remarkably, measure the Hubble rate during inflation. Since potential energy usually dominates kinetic energy in inflationary models, a measure of $H$ would be tantamount to measuring the potential $V$, again quite remarkable in view of the possibility that inflation was generated by physics at energy scales above $10^{15}$ GeV, 11 orders of magnitude beyond the capacity of present-day accelerators. There is no guarantee that gravitational waves produced during inflation will be detectable. Indeed, since $H^2 \propto \rho/m_{\rm Pl}^2$, where $m_{\rm Pl}^2 = 1/G$, the power spectrum is proportional to $\rho/m_{\rm Pl}^4$, the energy density at the time of inflation in units of the Planck mass. If inflation takes place at scales sufficiently smaller than the Planck scale, then primordial gravitational waves will not be detectable. Later in the book, we will develop the machinery necessary to answer the question, How small can the gravitational-wave amplitude be and still be detected?

Although we have not shown this fact, the fluctuations in $h$ are very close to Gaussian, just as are the quantum-mechanical fluctuations of the simple harmonic oscillator, and more generally, of any approximately free quantum field. Very near Gaussianity of perturbations is a robust prediction of inflation, which has been confirmed in the CMB and large-scale structure, both of which have placed increasingly tight upper limits on primordial non-Gaussianity. That said, some level of non-Gaussianity in particular of scalar perturbations is expected even in inflation. If detected, it would open another window into inflationary physics.

## 7.4 Scalar perturbations

One of the main goals of this chapter is to find the spectrum of scalar perturbations emerging from inflation. In principle, we need to specify the initial density and velocity perturba-

tions for each species. Fortunately, one of the primary predictions of single-field inflation is that it generates *adiabatic perturbations*: different patches of the universe have different overdensities, but the fractional density perturbations are the same for all species:

$$\frac{\delta\rho_s}{\rho_s} = \frac{\delta\rho}{\rho}, \tag{7.43}$$

with analogous relations for their velocities. The fundamental reason for this is the fact that inflation is driven by a single field, whose value determines when inflation ends. Thus, any given patch during inflation is completely characterized by the value of the field $\phi(\boldsymbol{x}, t)$. The adiabatic nature of perturbations has been confirmed to great precision by the CMB, which allows for different primordial density perturbations in the different species (referred to as *isocurvature* perturbations) at most as a percent-level fraction of the adiabatic perturbations.

This prediction also simplifies our task, since it is sufficient to derive $\delta\rho$. Equivalently, using the Einstein equations, we can specify the initial conditions in terms of $\Psi$, since specifying a single initial field is sufficient (in Sect. 7.5 we will derive how the density perturbations for all species follow from $\Psi$; also, $\Phi$ is simply equal and opposite to $\Psi$ in the regime of interest). The computation of $\Psi$ turns out to be more complicated than the tensor case considered earlier. The primary complication is the presence of perturbations in the scalar field $\phi$ that couple to $\Psi$. It will then all come down to determining how the perturbation to the scalar field, $\delta\phi$, is converted to $\Psi$.

Instead of dealing with the problem of mixing between $\Psi$ and the scalar field, we will first ignore it: in Sect. 7.4.1, we compute the spectrum of perturbations in the scalar field $\phi$ generated during inflation, neglecting $\Psi$. This turns out to be relatively simple to do, since it is virtually identical to the tensor calculation we went through above. Why are we justified in neglecting $\Psi$ and how do the perturbations get transferred from $\phi$ to $\Psi$? The next two subsections take turns answering this question from two different points of view. First, Sect. 7.4.2 argues that—in a sense to be defined there—until a mode moves outside the horizon, $\Psi$ is indeed negligibly small. Once it is far outside the horizon, this no longer holds, but we will find that a linear combination of $\Psi$ and $\delta\phi$ is conserved. This will allow us to convert the initial spectrum for $\delta\phi$ into a final spectrum for $\Psi$. The second way of justifying the initial neglect of $\Psi$ is to switch gauges and work in a gauge in which the spatial part of the metric is unperturbed, a so-called *spatially flat slicing*. In such a gauge, the calculation of Sect. 7.4.1 is exact; the only question remaining is how to convert back to conformal-Newtonian gauge to move on with the rest of the book. In Sect. 7.4.3, we solve this via one of the *gauge-invariant variables* we discussed in Sect. 6.2, which do not change upon a gauge transformation. First, we identify one which is proportional to $\delta\phi$ in a spatially flat slicing. It is then a simple matter to determine this variable in conformal Newtonian gauge, thereby linking $\Psi$ in conformal-Newtonian gauge to $\delta\phi$ in spatially flat slicing. Note that the two solutions to the coupling problem, as worked out in Sect. 7.4.2 and Sect. 7.4.3, are simply alternative approaches to the same problem. If you are comfortable with gauge transformations (working through Sect. 6.2 and the associated exercises

should get you to that point), Sect. 7.4.3 is a more elegant and direct approach; the more brute-force approach of Sect. 7.4.2 gives the same answer though and requires less formalism and background.

### 7.4.1 Scalar field perturbations around an unperturbed background

Let us decompose the scalar field into a zeroth-order homogeneous part and a perturbation,

$$\phi(\boldsymbol{x}, t) = \bar{\phi}(t) + \delta\phi(\boldsymbol{x}, t), \tag{7.44}$$

where now we distinguish the background field with an overbar. We want to find an equation governing $\delta\phi$ in a smooth expanding universe, i.e., with metric $g_{00} = -1$; $g_{ij} = \delta_{ij}a^2(\eta)$ (in the language of inflation practitioners, we are "ignoring gravity").

Consider again the conservation of the energy-momentum tensor, Eq. (7.11). The $\nu = 0$ component of this equation gives the desired equation for $\delta\phi$, although we now have to expand out to first order. Since we are assuming an unperturbed metric, the only first-order pieces are perturbations in the energy-momentum tensor. All the $\Gamma$ are either of zeroth order ($\Gamma^0{}_{ij} = \delta_{ij}a^2H$ and $\Gamma^i{}_{0j} = \Gamma^i{}_{j0} = \delta_{ij}H$) or zero (the rest of the components), as we found in Eqs. (2.24)–(2.25). So, writing the perturbed part of the energy-momentum tensor as $\delta T^\mu{}_\nu$ and considering the $\nu = 0$ component of the perturbed conservation equation leads to

$$0 = \frac{\partial}{\partial t}\delta T^0{}_0 + ik_i\delta T^i{}_0 + 3H\delta T^0{}_0 - H\delta T^i{}_i. \tag{7.45}$$

Our next task is to determine the perturbations to the energy-momentum tensor in terms of the perturbations to the scalar field.

First let us compute $\delta T^i{}_0$. Since the time-space components of the metric are zero, the second set of terms in Eq. (7.6), those with prefactor $g^{\alpha\beta}$, vanish. Therefore,

$$T^i{}_0 = g^{i\nu}\phi_{,\nu}\phi_{,0} \tag{7.46}$$

where we have returned to using $_{,\nu}$ to denote the derivative with respect to $x^\nu$. Since $g^{i\nu} = a^{-2}\delta_{i\nu}$, the index $\nu$ must be equal to $i$. Recall that the zeroth-order field $\bar{\phi}$ is homogeneous, so $\bar{\phi}_{,i} = 0$. The time-space component of the energy-momentum tensor therefore has no zeroth-order piece. To extract the first-order piece, we can set $\phi_{,i}$ to $\delta\phi_{,i} \to ik_i\delta\phi$. Then, setting all other factors to their zeroth-order values leads to

$$\delta T^i{}_0 = \frac{ik_i}{a^3}\bar{\phi}'\delta\phi. \tag{7.47}$$

The additional factor of $a$ enters the denominator here because $\bar{\phi}_{,0} = \bar{\phi}'/a$ (recall that the coordinate derivative is with respect to $t$).

The time-time component of the energy-momentum tensor is hardly any more difficult:

$$T^0{}_0 = g^{00}(\phi_{,0})^2 - \frac{1}{2}g^{\alpha\beta}\phi_{,\alpha}\phi_{,\beta} - V. \tag{7.48}$$

Setting $\phi(\boldsymbol{x},t)=\bar{\phi}(t)+\delta\phi(\boldsymbol{x},t)$ leads to

$$T^0{}_0=-\frac{1}{2}\left(\bar{\phi}_{,0}+\delta\phi_{,0}\right)^2-\frac{1}{2a^2}\delta\phi_{,i}\delta\phi_{,i}-V(\bar{\phi}+\delta\phi). \tag{7.49}$$

The spatial derivatives come in pairs, and pairs of first-order variables ($\delta\phi_{,i}$) lead to second-order terms. These may therefore be neglected. The potential may be expanded as a zeroth-order term, $V(\bar{\phi})$ plus a first-order correction, $V_{,\phi}\delta\phi$, where $V_{,\phi}$ is always evaluated at $\bar{\phi}$, so the first-order correction to the energy-momentum tensor is

$$\begin{aligned}\delta T^0{}_0&=-\bar{\phi}_{,0}\delta\phi_{,0}-V_{,\phi}\delta\phi\\&=-\frac{\bar{\phi}'\delta\phi'}{a^2}-V_{,\phi}\delta\phi.\end{aligned} \tag{7.50}$$

Similarly, you can show that the space-space component is

$$\delta T^i{}_j=\delta_{ij}\left(\frac{\bar{\phi}'\delta\phi'}{a^2}-V_{,\phi}\delta\phi\right). \tag{7.51}$$

Therefore, the conservation equation (7.45) becomes

$$\left(\frac{1}{a}\frac{\partial}{\partial\eta}+3H\right)\left(\frac{-\bar{\phi}'\delta\phi'}{a^2}-V_{,\phi}\delta\phi\right)-\frac{k^2}{a^3}\bar{\phi}'\delta\phi-3H\left(\frac{\bar{\phi}'\delta\phi'}{a^2}-V_{,\phi}\delta\phi\right)=0. \tag{7.52}$$

Carrying out the time derivatives (the only subtle one is $\partial V_{,\phi}/\partial\eta=V_{,\phi\phi}\bar{\phi}'$), multiplying by $a^3$, and collecting terms leads to

$$-\bar{\phi}'\delta\phi''+\delta\phi'\left(-\bar{\phi}''-4aH\bar{\phi}'-a^2V_{,\phi}\right)+\delta\phi\left(-a^2V_{,\phi\phi}\bar{\phi}'-k^2\bar{\phi}'\right)=0. \tag{7.53}$$

The $V_{,\phi\phi}$ term here is typically small, proportional to the slow-roll variables $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$ (Exercise 7.8), so it can be neglected. The coefficient of $\delta\phi'$, the first set of parentheses, is equal to $-2aH\bar{\phi}'$ using the zeroth-order equation (7.15), so after dividing by $-\bar{\phi}'$, we are left with

$$\delta\phi''+2aH\delta\phi'+k^2\delta\phi=0. \tag{7.54}$$

This equation for perturbations to $\delta\phi$ is identical to Eq. (7.27) for tensor perturbations to the metric: by neglecting $V_{,\phi\phi}$, we have essentially set the mass of the inflaton to zero, so $\delta\phi$ obeys the equation of a massless field in an expanding universe just like the massless gravitons. Thus we can trivially copy our result from Sect. 7.3.2 and immediately write that the power spectrum of fluctuations in $\delta\phi$ is equal to

$$P_{\delta\phi}=\frac{H^2}{2k^3}. \tag{7.55}$$

Compare this with Eq. (7.42). It is identical apart from a factor of $16\pi$G. Recall that we inserted this factor in the tensor case (with a bit of hand-waving; see Sect. 7.3.2) to turn

the dimensionless $h$ into a field with dimensions of mass. To get the result for $\delta\phi$, which is already a scalar field with the proper dimensions, we just remove this factor.

Before moving on, notice that the spatial part of the stress-energy tensor Eq. (7.51) is diagonal, i.e. proportional to $\delta^i{}_j$. We know from Ch. 6 that this means that there is no anisotropic stress, and that the Einstein equations imply $\Phi = -\Psi$. This makes our life simpler, since we can work with only one of the gravitational potentials. We will choose $\Psi$.

### 7.4.2 Super-horizon perturbations

Until now, we have neglected the metric perturbations. When the wavelength of the perturbation is much smaller than the horizon, this approximation is valid, as we will shortly see. In the process of seeing this, we will also find that, by the end of inflation, the metric perturbation has become important. So, although the inflation-induced perturbations start out all-"$\delta\phi$," they end up as a linear combination of $\Psi$ and $\delta\phi$ or more generally as a linear combination of $\Psi$ and perturbations to the energy-momentum tensor. The trick is to find the linear combination that is conserved outside the horizon. The value of this conserved linear combination is determined by $\delta\phi$ at horizon crossing (see the discussion after Eq. (7.42)); we can then evaluate it after inflation solely in terms of $\Psi$. The resulting equation will be of the form $\Psi \propto \delta\phi$ with the left-hand side being the post-inflation metric perturbation and the right the scalar field perturbation at horizon crossing. We can then finally relate our desired $P_\Psi$ to the $P_{\delta\phi}$ of Eq. (7.55).

Let us begin by rewriting the equation for conservation of energy, this time in the presence of metric perturbations. It is straightforward to show that Eq. (7.45) gets generalized to

$$\frac{\partial}{\partial t}\delta T^0{}_0 + ik_i\delta T^i{}_0 + 3H\delta T^0{}_0 - H\delta T^i{}_i + 3(\rho+\mathcal{P})\dot{\Psi} = 0 \tag{7.56}$$

where $\mathcal{P}$ and $\rho$ are the zeroth-order pressure and energy density, and we have replaced $\dot{\Phi}$ with $-\dot{\Psi}$. Were we correct to neglect $\Psi$ in the previous section?

We were, as long as the last term is significantly smaller than the others, which is true during inflation. The Einstein equations yield $\Psi \sim \delta T^0{}_0/\rho$, as we will verify shortly. This means that all terms in Eq. (7.56) except the last are of order $\rho\Psi$. On the other hand, one of the conditions of slow-roll inflation is that $|\rho+\mathcal{P}| \ll \rho$. In terms of the slow-roll parameters, $(\rho+\mathcal{P})/\rho \simeq 2\epsilon_{\rm sr}/3$. Thus, the last term in Eq. (7.11) is suppressed.

The above argument holds only during inflation. We now need to follow the perturbations as they exit the horizon, and then track their super-horizon evolution through the end of inflation. It is inevitable that the inequality $|\rho+\mathcal{P}| \ll \rho$ will break down sometime before the end of inflation. More physically speaking, at some point we need to convert the perturbations in the inflaton field, which decays into Standard Model particles most likely through some long complicated chain of reactions, into those in the gravitational potential. We already expect that, for adiabatic perturbations, the latter is all that counts.

One way to deal with the coupling between the metric perturbations and those of the energy density is to define the *curvature perturbation* $\mathcal{R}$ (we will understand where this

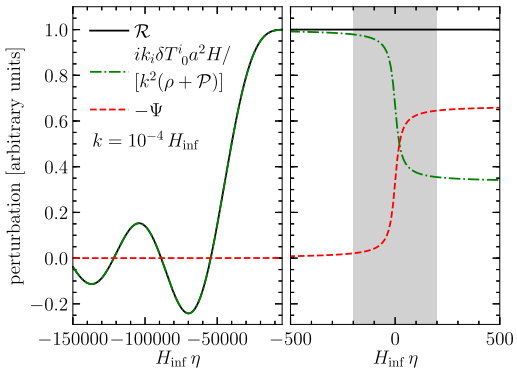



**FIGURE 7.6** Evolution of the curvature perturbation $\mathcal{R}$ during and after inflation ($H_{\rm inf}$ denotes the Hubble rate during slow-roll inflation). During inflation, $\mathcal{R}$ oscillates before freezing out when $k\eta \simeq -1$, at which point it leaves the horizon (left panel). Also shown are its constituents (right-hand side of Eq. (7.57)). During inflation, only the first term is relevant (dash-dotted line), leading to Eq. (7.58), while $\Psi$ is negligible (dashed curve). This changes when inflation concludes (right panel) and reheating happens (grey shaded area; notice the different scales of the $x$ axes in both panels). The evolution of the dot-dashed and dashed curves during that epoch depends on the microphysical model. However, $\mathcal{R}$ remains constant outside the horizon throughout this epoch, and we know how it is related to $\Psi$ once radiation domination takes over. Thus, we do not need to know what happens in the shaded area.

name comes from shortly):

$$\mathcal{R}(\boldsymbol{k},\eta) \equiv \frac{ik_i\delta T^i{}_0(\boldsymbol{k},\eta)a^2H(\eta)}{k^2[\rho+\mathcal{P}](\eta)} - \Psi(\boldsymbol{k},\eta). \tag{7.57}$$

We know that during inflation $\Psi$ is negligible compared to the first term. Further, $\rho+\mathcal{P} = (\bar{\phi}'/a)^2$ from Eq. (7.8) and Eq. (7.9); and Eq. (7.47) fixes the numerator of the first term in $\mathcal{R}$. We are left with

$$\mathcal{R} = -\frac{aH}{\bar{\phi}'}\delta\phi \qquad \text{(during inflation)}. \tag{7.58}$$

After inflation ends, during radiation domination,[3] $ik_i\delta T^i{}_0 = -4k\rho_{\rm r}\Theta_1/a$, proportional to the dipole of the radiation (which follows from Eq. (3.86); see also Sect. 7.5 below). Since the pressure of radiation is equal to a third of the energy density,

$$\begin{aligned}\mathcal{R} &= -\frac{3aH\Theta_1}{k} - \Psi \\ &= -\frac{3}{2}\Psi \qquad \text{(post inflation, radiation domination)}.\end{aligned} \tag{7.59}$$

We will derive the second equality in Sect. 7.5 below. A sketch of how $\mathcal{R}$, $\Psi$, and $k_i\delta T^i{}_0$ evolve during and after inflation is shown in Fig. 7.6.

[3] So to be precise, we here assume that we wait long enough after the end of inflation so that we are safely in the radiation-dominated epoch.

The variable $\mathcal{R}$ is so important because it is conserved when the perturbation moves outside the horizon (from $H_{\rm inf}\eta \gtrsim -10^4$ onwards in Fig. 7.6). We will prove this conservation shortly, but first let us appreciate its importance. Since we know that, after inflation, $\mathcal{R} = -3\Psi/2$, we can immediately relate $\Psi$ coming out of inflation to $\delta\phi$ at horizon crossing,

$$\Psi\Big|_{\text{post inflation}} = \frac{2}{3} aH \frac{\delta\phi}{\bar{\phi}'}\Big|_{\text{horizon crossing}} . \tag{7.60}$$

Equivalently, the post-inflation power spectrum of $\Psi$ is simply related to the horizon-crossing spectrum of $\delta\phi$:

$$\begin{aligned} P_\Psi(k)\Big|_{\text{post inflation}} &= \frac{4}{9}\left(\frac{aH}{\bar{\phi}'}\right)^2 P_{\delta\phi}(k)\Big|_{aH=k} \\ &= \frac{2}{9k^3}\left(\frac{aH^2}{\bar{\phi}'}\right)^2\Big|_{aH=k} , \end{aligned} \tag{7.61}$$

the second line following from Eq. (7.55). Another way to express the power spectrum of scalar perturbations is to eliminate $\bar{\phi}'$ in favor of the slow-roll parameter $\epsilon_{\rm sr}$. You will show (Exercise 7.7) that $(aH/\bar{\phi}')^2 = 4\pi G/\epsilon_{\rm sr}$, so

$$P_\Psi(k) = P_\Phi(k) = \frac{8\pi G}{9k^3}\frac{H^2}{\epsilon_{\rm sr}}\Big|_{aH=k} . \tag{7.62}$$

The first equality here follows from our ubiquitous assumption that anisotropic stresses are small, so that $\Psi = -\Phi$. Comparing to Eq. (7.42), we see that the ratio of scalar to tensor modes is of order $1/\epsilon_{\rm sr}$; that is, we expect scalar perturbations to dominate. Finally, another way of writing the scalar power spectrum is to eliminate $\epsilon_{\rm sr}$ in favor of the inflaton potential and its derivative, using the result of Exercise 7.8,

$$P_\Phi(k) = P_\Psi(k) = \frac{128\pi^2 G^2}{9k^3}\left(\frac{HV}{V_{,\phi}}\right)^2\Big|_{aH=k} . \tag{7.63}$$

This equation allows for some important insights into the generation of scalar perturbations during inflation. As we have seen in the previous section, the amplitude of the perturbations to $\phi$ depends only on the Hubble rate during inflation, as is true for the tensor modes. But in the end we care about the gravitational potentials, and have found that their amplitude depends on the slope of the scalar field potential. In particular, by making $V_{,\phi}/V$ sufficiently small, we can counteract a smaller Hubble rate during inflation while keeping $P_\Phi(k)$ fixed; importantly, this then lowers the amplitude of tensor perturbations relative to scalar ones. Where does this dependence come from?

A physical way to think about this is to recall that $\Phi$ quantifies the perturbation to the scale factor, $\Phi = \delta a/a$, where $\delta a$ in a given region is positive if that region has expanded more than the average region during inflation. This perturbation to the scale factor is related to a small change in cosmic time via $\delta a = \dot{a}\delta t = aH\delta t$. But, we can also express the

clock change $\delta t$ in terms of a difference in the value of the scalar field through $\delta\phi = \dot{\phi}\delta t$; in other words, the value of the field $\phi$ (which, unlike $t$, is a coordinate invariant) provides the physical clock during inflation. Equating these two expressions for $\delta t$, one in terms of $\Phi$ and the other in terms of $\delta\phi$, we find

$$\Phi \sim H\frac{\delta\phi}{\dot{\phi}}. \tag{7.64}$$

Indeed, up to a factor of order 1, the conversion factor in the first line of Eq. (7.61) is precisely the square of this result. Therefore, the power spectrum of the gravitational potential is inversely proportional to $\dot{\phi}^2$. But, the equations of motion for a slowly rolling scalar field, neglecting the first term in (7.15), dictate that the field travels more rapidly in a steeper potential; that is, $\dot{\phi} \propto V_{,\phi}$. Therefore, the power spectrum of $\Phi$ is larger if the potential is shallower.

It now remains to prove that $\mathcal{R}$ is conserved on super-horizon scales. To see this, let us turn to the conservation equation (7.56). On large scales, $k_i\delta T^i{}_0$ is proportional to $k^2$ and so can be ignored, leaving

$$\frac{\partial}{\partial t}\delta T^0{}_0 + 3H\delta T^0{}_0 - H\delta T^i{}_i = -3(\rho+\mathcal{P})\dot{\Psi} \qquad \text{(super-horizon)}. \tag{7.65}$$

On large scales, you will show (Exercise 7.13) that the energy-momentum tensor satisfies

$$\frac{ik_i\delta T^i{}_0 a^2 H}{k^2} = -\frac{\delta T^0{}_0}{3}. \tag{7.66}$$

Therefore, on large scales

$$\mathcal{R} = -\Psi - \frac{1}{3}\frac{\delta T^0{}_0}{\rho+\mathcal{P}}. \tag{7.67}$$

Eliminating $\Psi$ in favor of $\mathcal{R}$ in the conservation equation leads to

$$\frac{\partial}{\partial t}\delta T^0{}_0 + 3H\delta T^0{}_0 - H\delta T^i{}_i = 3(\rho+\mathcal{P})\frac{\partial\mathcal{R}}{\partial t} + (\rho+\mathcal{P})\frac{\partial}{\partial t}\left[\frac{\delta T^0{}_0}{\rho+\mathcal{P}}\right]. \tag{7.68}$$

The partial derivative on the right acting on $\delta T^0{}_0$ cancels the first term on the left, leaving

$$\delta T^0{}_0\left[3H + \frac{1}{\rho+\mathcal{P}}\left(\frac{d\rho}{dt} + \frac{d\mathcal{P}}{dt}\right)\right] - H\delta T^i{}_i = 3(\rho+\mathcal{P})\frac{\partial\mathcal{R}}{\partial t}. \tag{7.69}$$

Recall from Eq. (2.56) that $d\rho/dt = -3H(\rho+\mathcal{P})$, so we can rewrite the left-hand side as

$$3H\left[\frac{\dot{\mathcal{P}}}{\dot{\rho}}\delta\rho - \delta\mathcal{P}\right] = 3(\rho+\mathcal{P})\frac{\partial\mathcal{R}}{\partial t}, \tag{7.70}$$

since $-\delta T^0{}_0$ is the perturbation to the energy density, while $\delta T^i{}_i/3$ is the perturbation to the pressure. Thus, $\partial\mathcal{R}/\partial t = 0$ precisely if

$$\delta\mathcal{P} = \frac{\dot{\mathcal{P}}}{\dot{\rho}}\delta\rho. \tag{7.71}$$

At the background level, we can write $\dot{\mathcal{P}} = (d\mathcal{P}/d\bar{\phi})\dot{\bar{\phi}}$, and likewise for $\dot{\rho}$, since $\bar{\phi}$ is our clock and equivalent to the time coordinate. For a single scalar field, both $\mathcal{P}$ and $\rho$ are unique functions of the field $\phi$ (see Eq. (7.8) and Eq. (7.9)), so we can also write $\delta\mathcal{P} = (d\mathcal{P}/d\bar{\phi})\delta\phi$, and correspondingly for $\delta\rho$. Combining these facts proves that Eq. (7.71) holds in single-field inflation. Thus, in this case $\mathcal{R}$ is indeed conserved on large scales. This changes in more complicated inflationary models where several fields are active, and where $\mathcal{R}$ in general evolves outside the horizon.

### 7.4.3 Spatially flat slicing

The treatment of the previous subsection is complete, but it is not the most elegant way to understand scalar perturbations in inflation. A much simpler way is to move back and forth between different gauges, making use along the way of the concept of a gauge-invariant variable introduced in Sect. 6.2. Here we outline this method, leaving some of the more detailed calculations as exercises.

We saw earlier that one of the major complications in conformal-Newtonian gauge was that perturbations to the scalar field $\delta\phi$ are coupled to the potential $\Psi$. It would be nice to transform to a gauge in which these perturbations decoupled. Consider a gauge with *spatially flat slicing*, such that the spatial part of the metric obeys $g_{ij} = a^2\delta_{ij}$. In this gauge the line element is

$$ds^2 = -[1 + 2A(\boldsymbol{x}, t)]dt^2 - 2a(t)B_{,i}(\boldsymbol{x}, t)dx^i dt + a^2(t)\delta_{ij}dx^i dx^j, \tag{7.72}$$

i.e., there are two functions $A$ and $B$ characterizing the scalar perturbations. In this case, the equation for $\delta\phi$ is given exactly by Eq. (7.54): the perturbations in the scalar field do not couple to $A$, $B$. Therefore, without having to neglect any couplings, we can identify the power spectrum for $\delta\phi$ as given by Eq. (7.55).

The next step is to chose a gauge-invariant variable, one that remains the same when transforming from one gauge to the next. Bardeen (1980) identified several such variables, two characterizing scalar perturbations to the metric and two characterizing perturbations to the matter. Any linear combination of these is still gauge invariant. We would like to identify the combination that is proportional to $\delta\phi$ in the gauge with spatially flat slicing.

One of Bardeen's variables, which we shall call $\mathcal{V}$, is particularly useful. It is defined as

$$\mathcal{V}(\boldsymbol{k}, t) \equiv B(\boldsymbol{k}, t) + \frac{ik_i}{k^2}\frac{a\delta T^i{}_0(\boldsymbol{k}, t)}{\rho + \mathcal{P}}. \tag{7.73}$$

In conformal-Newtonian gauge, for matter, $\mathcal{V}$ is directly related to the velocity through $\boldsymbol{u}_m = i\boldsymbol{k}\mathcal{V}$. For radiation in the same gauge, $ik\mathcal{V} = -3i\Theta_{\mathrm{r},1}$, i.e. it is proportional to the

dipole. In the spatially flat gauge, Eq. (7.73) is

$$\mathcal{V} = B - \frac{\bar{\phi}'\delta\phi}{(\rho+\mathcal{P})a^2} \qquad \text{(spatially flat slicing)} \tag{7.74}$$

where we have evaluated $\delta T^i{}_0$ with Eq. (7.47). Now, Bardeen's $\Phi_H$ (Eq. (6.19)) in the spatially flat gauge is equal to $aHB$ since $D = E = 0$ in this gauge. Since both $\Phi_H$ and $\mathcal{V}$ are gauge invariant, the combination

$$\mathcal{R} \equiv -\Phi_H + aH\mathcal{V} \tag{7.75}$$

is also gauge invariant. In spatially flat slicing, it is equal to

$$\mathcal{R} = -\frac{aH}{\bar{\phi}'}\delta\phi \qquad \text{(spatially flat slicing)}. \tag{7.76}$$

We can thus immediately relate the power in $\mathcal{R}$ to the power in $\delta\phi$,

$$P_{\mathcal{R}}(k) = \left(\frac{aH}{\bar{\phi}'}\right)^2 P_{\delta\phi}(k). \tag{7.77}$$

We know $P_{\delta\phi}$ from Eq. (7.55), and the prefactor is $4\pi G/\epsilon_{\rm sr}$, so

$$P_{\mathcal{R}}(k) = \left.\frac{2\pi G H^2}{\epsilon_{\rm sr} k^3}\right|_{aH=k}. \tag{7.78}$$

Eq. (7.78) is very useful, for it expresses the power spectrum of a gauge-invariant quantity. For this reason, constraints on the power spectrum of primordial scalar perturbations are usually phrased in terms of $P_{\mathcal{R}}(k)$. Although we computed it in the spatially-flat coordinates of Eq. (7.72), once we have this answer, we can compute $\mathcal{R}$ in any gauge and then relate the power in the perturbation variables of that gauge to $P_{\mathcal{R}}$.

Throughout this book, we have been working in conformal-Newtonian gauge. In this gauge, $\Phi_H = -\Phi$, so $\mathcal{R}$ as defined in Eq. (7.75) is indeed given by Eq. (7.57), noting that $B = 0$ in this gauge. We argued in Sect. 7.4.2 that in conformal-Newtonian gauge, after inflation, $\mathcal{R} = 3\Phi/2$, so $P_\Phi = 4P_{\mathcal{R}}/9$, or using Eq. (7.78),

$$P_\Phi(k) = \left.\frac{8\pi G H^2}{9k^3\epsilon_{\rm sr}}\right|_{aH=k}, \tag{7.79}$$

in exact agreement with our earlier calculation, Eq. (7.62).

For us, this is the end of the calculation of inflationary perturbations, but not quite the end of the story. Bardeen and others have argued that $\Phi_H$ has a nice geometrical interpretation, one shared by $\mathcal{R}$ in certain gauges. In particular, the curvature of the three-dimensional space at fixed time is equal to $4k^2\Phi_H/a^2$ (Exercise 3.13). Therefore, perturbations in $\Phi_H$ represent *curvature perturbations*: even though the zeroth-order space is Euclidean, perturbations induce a curvature that varies from place to place. In conformal-Newtonian gauge or in a spatially-flat slicing this interpretation would seem irrelevant to

perturbations in $\mathcal{R}$, since $\mathcal{R}$ is a combination of both $\Phi_H$ and the velocity. However, if one moves to a *comoving gauge*, one in which the velocities vanish, then $\mathcal{R}$ is equal to $\Phi_H$. In comoving gauges, then, it is clear that $\mathcal{R}$ corresponds to a curvature perturbation, and indeed the scalar perturbations generated during inflation are often called curvature perturbations. The reason $\mathcal{R}$ is so ubiquitous is that it is simply the most convenient way to describe the adiabatic perturbations from inflation.

## 7.5 The Einstein–Boltzmann equations at early times

What is remaining now is just one final step to connect the perturbation $\Psi$ to the nine perturbation variables we need to track following the equations we derived in Ch. 5 and Ch. 6. In principle, we need initial conditions for all of these variables; fortunately, we now know that perturbations from inflation are adiabatic (Eq. (7.43)), and the initial conditions become very simple in that case. Since the fractional perturbations in all species are the same, they are all determined by $\Psi$, and we only need to work out how this relation looks at early times. In addition, in the previous sections we assumed some relations between $\Psi$, $\Phi$ and $\Theta_1$ which we shall now prove along the way.

Let us consider first the Boltzmann equations (5.67)–(5.73) at very early times after the end of inflation (i.e. $\eta > 0$ but small). In particular, we want to consider times so early that, for any $k$-mode of interest, $k\eta \ll 1$ or equivalently $k/aH \ll 1$. This inequality immediately leads to several important simplifications. Consider Eq. (5.67):

$$\Theta' + ik\mu\Theta = -\Phi' - ik\mu\Psi - \tau'\left[\Theta_0 - \Theta + \mu u_{\rm b} - \frac{1}{2}\mathcal{P}_2(\mu)\Pi\right]. \tag{7.80}$$

The first term is of order $\Theta/\eta$, while the second is of order $k\Theta$. Therefore, the first is larger than the second by a factor of order $1/(k\eta)$, which, by assumption, is much greater than 1. In a similar way, we can argue that all terms in the Boltzmann equations multiplied by $k$ can be neglected at early times; this also applies to $u_{\rm b}$ and $\Pi$, as we will argue below. Physically, this means that, at early times, all perturbations of interest have wavelengths ($\sim k^{-1}$) much larger than the distance over which causal physics operates. A hypothetical observer then who sees only photons from within her causal horizon will see a uniform sky. Thus higher multipoles ($\Theta_1,\ \Theta_2,\ \ldots$) are much smaller than the monopole, $\Theta_0$. Therefore, the perturbations to the photon and neutrino temperatures evolve according to

$$\begin{aligned}\Theta_0' + \Phi' &= 0,\\ \mathcal{N}_0' + \Phi' &= 0.\end{aligned} \tag{7.81}$$

The same principles can be applied to the matter distributions. The overdensity equations reduce to

$$\begin{aligned}\delta_{\rm c}{}' &= -3\Phi',\\ \delta_{\rm b}{}' &= -3\Phi'.\end{aligned} \tag{7.82}$$

Outside the comoving horizon, gravity is the only relevant force. This is the reason that dark matter and baryons follow the same equation: gravity does not care whether it is acting on a dark matter particle or a proton. The velocities are smaller than the overdensities by a factor of order $k\eta$ and may be set to zero initially.

Now let us turn to the Einstein equations at early times. First consider Eq. (6.41). The first term there contains a factor of $k^2$ so may be neglected (notice that we are taking the opposite limit to the Newtonian, no-expansion approximation, where one uses $k \gg aH$). Also, the two matter terms on the right are negligible at early times since radiation dominates. Therefore, we have

$$3\frac{a'}{a}\left(\Phi' - \frac{a'}{a}\Psi\right) = 16\pi G a^2 \rho_{\rm r}\Theta_{\rm r,0}, \tag{7.83}$$

where the total radiation monopole is defined in Eq. (6.79). But since radiation dominates, $a \propto \eta$ (recall Eq. (2.95) and the discussion immediately afterward) so $a'/a = aH = 1/\eta$. Therefore,

$$\begin{aligned}\frac{\Phi'}{\eta} - \frac{\Psi}{\eta^2} &= \frac{16\pi G\rho a^2}{3}\Theta_{\rm r,0} \\ &= \frac{2}{\eta^2}\Theta_{\rm r,0}\end{aligned} \tag{7.84}$$

where the last equality follows by virtue of the zeroth-order Einstein (i.e., Friedmann) equation. Then, multiplying Eq. (7.84) by $\eta^2$ leads to

$$\Phi'\eta - \Psi = 2\Theta_{\rm r,0}. \tag{7.85}$$

Recall that Eq. (7.81) relates the time derivative of the monopoles to the derivative of the potential. We can therefore eliminate both monopoles from Eq. (7.85) by differentiating both right- and left-hand sides. Then,

$$\Phi''\eta + \Phi' - \Psi' = -2\Phi' \tag{7.86}$$

where the right-hand side follows since both $\Theta_0'$ and $\mathcal{N}_0'$ are equal to $-\Phi'$ for these large-scale modes.

So far we have used only one Einstein equation. The second, Eq. (6.48), describes how the higher moments of the photon and neutrino distributions source $\Psi + \Phi$. Let us here neglect these higher-order moments, which lead to a slightly nonzero sum of the gravitational potentials.[4] Under this approximation, we can eliminate $\Psi$ everywhere by simply setting it to $-\Phi$. Then,

$$\Phi''\eta + 4\Phi' = 0. \tag{7.87}$$

[4] See Exercise 7.16 for a careful accounting of the effect of the neutrino quadrupole; the photon quadrupole is kept minuscule by Compton scattering, so it really does not contribute to Eq. (6.48).

Inserting the ansatz $\Phi = \eta^p$ leads to the algebraic equation

$$p(p-1) + 4p = 0, \tag{7.88}$$

which allows for two solutions: $p = 0$ and $p = -3$. The $p = -3$ mode is the decaying mode. If it is excited very early on, it will quickly die out and have no impact on the universe. The $p = 0$ mode, on the other hand, does not decay if excited. It is the mode we are interested in.

Focusing therefore on only the $p = 0$ mode, we see that Eq. (7.85) relates the gravitational potential to the total radiation overdensity:

$$\Phi = 2\Theta_{\mathrm{r},0}. \tag{7.89}$$

Thus, $\Theta_{\mathrm{r},0}$ as well as its constituents $\Theta_0$ and $\mathcal{N}_0$ remain constant in time. For adiabatic perturbations, Eq. (7.43) implies

$$\Theta_0(\boldsymbol{k}, \eta_i) = \mathcal{N}_0(\boldsymbol{k}, \eta_i), \tag{7.90}$$

which leads to

$$\Phi(\boldsymbol{k}, \eta_i) = 2\Theta_0(\boldsymbol{k}, \eta_i) \tag{7.91}$$

where we have explicitly written the $k$-dependence of all these variables and the fact that we are setting up the initial conditions at some early time $\eta_i$.

The initial conditions for matter, both $\delta_{\mathrm{c}}$ and $\delta_{\mathrm{b}}$, also become simple once we restrict to adiabatic perturbations. Combining the first relation of Eq. (7.81) and (7.82) leads to

$$\delta_{\mathrm{c}}(\boldsymbol{k}, \eta) = 3\Theta_0(\boldsymbol{k}, \eta) + \text{constant}(\boldsymbol{k}) \tag{7.92}$$

for the dark matter overdensity, with an identical equation for the baryon overdensity. Clearly, adiabatic perturbations require the same constant for CDM and baryons. We now show that this constant is in fact zero. Adiabatic perturbations must have a uniform matter-to-radiation ratio. This ratio is given by

$$\frac{n_{\mathrm{c}}}{n_\gamma} = \frac{\bar{n}_{\mathrm{c}}}{\bar{n}_\gamma}\left[\frac{1+\delta_{\mathrm{c}}}{1+3\Theta_0}\right]. \tag{7.93}$$

The prefactor, the ratio of zeroth-order number densities, is a constant in both space and time. For the ratio of matter-to-radiation number density to be uniform, therefore, the combination inside the brackets, which linearizes to $1 + \delta_{\mathrm{c}} - 3\Theta_0$, must be independent of space. So the perturbations must sum to zero, and we have

$$\delta_{\mathrm{c}} = \delta_{\mathrm{b}} = 3\Theta_0. \tag{7.94}$$

We also need the initial conditions for the velocities and dipole moments of matter and radiation, respectively. You will show in Exercise 7.17 that the appropriate initial conditions

are

$$\Theta_1(\boldsymbol{k},\eta) = \mathcal{N}_1(\boldsymbol{k},\eta) = \frac{iu_{\rm b}(\boldsymbol{k},\eta)}{3} = \frac{iu_{\rm c}(\boldsymbol{k},\eta)}{3} = -\frac{k}{6aH}\Phi(\boldsymbol{k},\eta). \tag{7.95}$$

With this, the equation for $\mathcal{R}$ on super-horizon scales during radiation domination, Eq. (7.59), simply becomes

$$\mathcal{R} = -\frac{3}{2}\Psi. \tag{7.96}$$

## 7.6 Summary

In order to understand how scales that should be uncorrelated today are observed to have almost identical temperatures, we are virtually forced into the theory of inflation. In addition to explaining away other nagging fine-tuning problems of the concordance cosmology such as the flatness problem (Exercise 7.1), inflation is also a mechanism for generating primordial perturbations around the smooth universe.

Inflation predicts that quantum-mechanical perturbations in the very early universe are first produced when the relevant scales are causally connected. Then these scales are whisked outside the horizon by inflation, only to re-enter much later to serve as initial conditions for the growth of structure in the universe. The perturbations are best described in terms of their Fourier modes. The mean of a given Fourier mode, for example for the gravitational potential, is zero:

$$\langle\Phi(\boldsymbol{k})\rangle = 0. \tag{7.97}$$

Further, any given Fourier mode is uncorrelated with a different one. However, a given mode has nonzero variance, so

$$\langle\Phi(\boldsymbol{k})\Phi^*(\boldsymbol{k}')\rangle = P_\Phi(k)(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}'), \tag{7.98}$$

the Dirac delta function enforcing the independence of the different modes. The property of approximate Gaussianity is equivalent to the statement that any higher-order correlation functions, for example involving three powers of $\Phi$, are highly suppressed.

In the case of scalar perturbations, the ones of most importance for us, the power spectrum is given by Eq. (7.62). Perturbations to the tensor part of the metric are also produced and are also Gaussian with mean zero; the power spectrum of tensor modes is given by Eq. (7.42). The scalar spectrum depends on the slow-roll parameter $\epsilon_{\rm sr}$, defined in Eq. (7.17), which is proportional to the derivative of the Hubble rate. Since the Hubble rate is close to constant during inflation—because of the dominance of potential energy—$\epsilon_{\rm sr}$ is typically small.

A spectrum in which $k^3P_\Phi(k)$ is constant (i.e., does not depend on $k$) is called a *scale-invariant* or *scale-free* spectrum. Apart from small deviations encoded in the slow-roll parameters, both the scalar and the tensor perturbations are scale-free. Moreover, since the

field rolls down a potential well during inflation, so that the Hubble rate slowly decreases, a further generic prediction is that the potential is slightly *red-tilted*, with larger-scale perturbations, those which left the horizon earlier, having a slightly larger amplitude than smaller-scale perturbations. A spectrum with a small red tilt has indeed been conclusively detected in the CMB.

The scalar perturbations generated during inflation are nowadays most commonly parametrized in terms of the power spectrum of the gauge-invariant curvature perturbation $\mathcal{R}$. This has the great advantage of being conserved on super-horizon scales, regardless of whether matter or radiation dominates, and is thus a good unambiguous anchoring point. From Eq. (7.78), we have

$$P_{\mathcal{R}}(k) = \frac{2\pi}{k^3}\frac{H^2}{m_{\rm Pl}^2\epsilon_{\rm sr}}\bigg|_{aH=k} \equiv 2\pi^2\mathcal{A}_s k^{-3}\left(\frac{k}{k_{\rm p}}\right)^{n_s-1}, \tag{7.99}$$

where $\mathcal{A}_s$ is the variance of curvature perturbations in a logarithmic wavenumber interval centered around the pivot scale $k_{\rm p}$, and $n_s$ is the scalar spectral index. The pivot scale is a matter of convention, and is often determined as the scale best constrained by a given set of observations (say, CMB anisotropies; the Planck team adopts $k_{\rm p} = 0.05$ Mpc$^{-1}$, and we do so as well). In our fiducial cosmology,

$$\mathcal{A}_s = \frac{k_{\rm p}^3}{2\pi^2}P_{\mathcal{R}}(k_{\rm p}) \simeq 2.1\times 10^{-9}. \tag{7.100}$$

Thus, the typical amplitude of curvature perturbations on the scale $k_{\rm p}$ is $\sqrt{\mathcal{A}_s} \simeq 4.6\times 10^{-5}$, which is of similar order of magnitude as (but a bit larger than) the temperature fluctuations in the CMB. We will see in Ch. 9 that this is no coincidence.

For tensor modes, we derived the power spectrum of a single polarization $P_h$ in Eq. (7.42). Primordial tensor modes are conventionally parametrized via their total power spectrum $P_{\rm T}(k)$ (outside the horizon) defined via

$$\left\langle h_{ij}^{\rm TT}(\boldsymbol{k})\left(h_{ij}^{\rm TT}\right)^*(\boldsymbol{k}')\right\rangle\Big|_{\eta=0} \equiv (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}')P_{\rm T}(k). \tag{7.101}$$

Performing the index summation via Eq. (6.49), the left-hand side evaluates to $2\langle h_+h_+^*\rangle + 2\langle h_\times h_\times^*\rangle$, so we have

$$P_{\rm T}(k) = 4P_h(k) = \frac{32\pi}{k^3}\frac{H^2}{m_{\rm Pl}^2}\bigg|_{aH=k} \equiv 2\pi^2\mathcal{A}_{\rm T}k^{-3}\left(\frac{k}{k_{\rm p}}\right)^{n_{\rm T}}, \tag{7.102}$$

which serves to define the conventional tensor amplitude $\mathcal{A}_{\rm T}$ and spectral index $n_{\rm T}$. Note that this convention—which has become common—says that a scale-free scalar spectrum corresponds to $n_s = 1$, while for the tensor modes the same statement is $n_{\rm T} = 0$. In practice, $\mathcal{A}_{\rm T}$ is often replaced with the *tensor-to-scalar ratio* $r$,

$$r(k) \equiv \frac{P_{\rm T}(k)}{P_{\mathcal{R}}(k)} \overset{k=k_{\rm p}}{=} \frac{\mathcal{A}_{\rm T}}{\mathcal{A}_s}. \tag{7.103}$$

Eq. (7.99) and Eq. (7.102) immediately yield

$$r(k) = 16\,\epsilon_{\rm sr}\Big|_{aH=k}. \tag{7.104}$$

While $r$ in principle depends on $k$ due to the different spectral index of scalar and tensor modes, this dependence is negligibly small in practical applications.

We can relate the primordial spectral indices $n_s$ and $n_{\rm T}$ to the slow-roll parameters $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$. Consider first the tensor spectrum. By virtue of the definition in Eq. (7.99),

$$\frac{d\ln P_{\rm T}(k)}{d\ln k} = n_{\rm T} - 3. \tag{7.105}$$

The logarithmic derivative has two terms, where the trivial one $d\ln(k^{-3})/d\ln k$ cancels the $-3$ here, leaving $n_{\rm T} = 2d\ln H/d\ln k$. The logarithmic derivative of the Hubble rate at horizon crossing is a bit subtle:

$$\frac{d\ln H}{d\ln k}\bigg|_{aH=k} = \frac{k}{H}\frac{dH}{d\eta} \times \frac{d\eta}{dk}\bigg|_{aH=k}. \tag{7.106}$$

By definition (Eq. (7.17)), $H' = -aH^2\epsilon_{\rm sr}$, and $d\eta|_{aH=k}/dk = -d(aH)^{-1}|_{aH=k}/dk = 1/k^2$, so

$$\frac{d\ln H}{d\ln k}\bigg|_{aH=k} = -\frac{k}{H}\frac{aH^2\epsilon_{\rm sr}}{k^2}\bigg|_{aH=k} = -\epsilon_{\rm sr}. \tag{7.107}$$

Therefore, the primordial spectral index of tensor perturbations produced by inflation is

$$n_{\rm T} = -2\epsilon_{\rm sr}. \tag{7.108}$$

The scalar spectral index follows from a similar argument. Taking the logarithmic derivative of $P_\Phi$ leads to

$$n_s - 1 = \frac{d}{d\ln k}[\ln(H^2) - \ln(\epsilon_{\rm sr})]. \tag{7.109}$$

The derivative of $H$ again gives $-2\epsilon_{\rm sr}$ while the logarithmic derivative of $\epsilon_{\rm sr}$ is $-2(\epsilon_{\rm sr} + \delta_{\rm sr})$ (Exercise 7.7). So,

$$n_s = 1 - 4\epsilon_{\rm sr} - 2\delta_{\rm sr}. \tag{7.110}$$

The fact that the tensor index $n_{\rm T}$ is proportional to $\epsilon_{\rm sr}$ leads to one of the robust predictions of inflation (Exercise 7.14). Many inflationary models have been proposed which offer different predictions for $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$. Almost all of these, however, maintain the feature that the ratio of tensor-to-scalar modes (which we saw earlier was proportional to $\epsilon_{\rm sr}$) is directly related to the tensor spectral index (here also seen to be directly proportional to $\epsilon_{\rm sr}$). As you progress through the remainder of this book, moving from the evolution of structure to its observational probes, try to bear in mind the crucial question of whether this prediction can be put to the observational test.

The slow-roll parameters are a convenient way to summarize the predictions of an inflationary model. However, ultimately we are interested in the physics, so we are interested in how these parameters relate back to the fundamental entity, the potential $V$ of the scalar field responsible for inflation. You will show in Exercise 7.8 that these parameters can be expressed in terms of the potential and its derivatives. Therefore, extracting the values of $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$ from the data is tantamount to probing the potential of the field driving inflation. Given that the scale of this potential could be on the order of $10^{15}$ GeV (Exercise 7.18), this is quite an impressive probe!

Inflation is a difficult subject to grasp fully, and it helps to go through the ideas several times via different angles. While everything we need for the remainder of the book is contained in the equations of this section, let us conclude this chapter with some suggested further reading. The initial article by Guth (1981) is completely accessible and as clear a statement possible of the problems that led to inflation and the initial attempt (old inflation) to solve them. The textbook *Physical Foundations of Cosmology* (Mukhanov, 2005) by one of the pioneers of the field provides a clear, modern overview of the physics of inflation and generation of perturbations. Finally, for those interested in the foundations of the generation of quantum fluctuations in the expanding universe, we recommend Birrell and Davies (1984).

## Exercises

**7.1** Inflation also solves the *flatness* problem. This is the question of why the energy density today is so close to critical.

**(a)** Suppose that

$$\Omega(t) \equiv \frac{8\pi G\rho(t)}{3H^2(t)} \tag{7.111}$$

is equal to 0.3 today, where $\rho$ counts the energy density in matter and radiation (ignore the cosmological constant). From Eq. (3.14), plot $\Omega(t) - 1$ as a function of the scale factor. How close to 1 would $\Omega(t)$ have been back at the Planck epoch (assuming no inflation took place so that the scale factor at the Planck epoch was of order $10^{-32}$)? This fine-tuning of the initial conditions is the flatness problem. If not for the fine tuning, an open universe would be *obviously* open (i.e., $\Omega$ would be almost exactly zero today); a closed universe would have recollapsed at very early times.

**(b)** Now show that inflation solves the flatness problem. Extrapolate $\Omega(t) - 1$ back to the end of inflation, and then through 60 $e$-folds of inflation. What is $\Omega(t) - 1$ right before these 60 $e$-folds of inflation? This shows how inflation "flattens" space (see Fig. 7.4).

**7.2** Another way of looking at the problems that inflation solves is to consider the entropy within our Hubble volume. This is proportional to the total number of particles in the volume, with a proportionality constant of order unity. How many photons are

there within our Hubble volume today? Explain how inflation produces entropy this large.

**7.3** We showed that, if the universe was always dominated by ordinary matter or radiation early on, then the comoving horizon when the scale factor was $a_e$ is given by $a_0 H_0 / a_e H_e$ times the comoving Hubble radius today. Compute this ratio assuming that the temperature was equal to $10^{14}$ GeV at $a_e$. Account for the radiation-to-matter transition at $a = a_{\rm eq}$.

**7.4** Derive the energy-momentum tensor for a canonical scalar field, whose Lagrangian is given by

$$\mathcal{L}_\phi = -\frac{1}{2} g^{\mu\nu} \frac{\partial \phi}{\partial x^\mu} \frac{\partial \phi}{\partial x^\nu} - V(\phi). \tag{7.112}$$

Recall that the Lagrangian is given by the difference of kinetic energy and potential energies, and that, since $g^{00} < 0$, we need a minus sign for the kinetic energy. The energy-momentum tensor is obtained by varying the action with respect to the metric:

$$T_{\mu\nu} = \frac{\delta \mathcal{L}_\phi}{\delta g^{\mu\nu}} + g_{\mu\nu} \mathcal{L}_\phi. \tag{7.113}$$

Use this to derive Eq. (7.6).

**7.5** Show that Eq. (7.15) follows from Eq. (7.14) by changing variables from $t$ to $\eta$.

**7.6** Consider a free, homogeneous scalar field with mass $m$. The potential for this field is $V = m^2 \phi^2 / 2$. Show that, if $m \gg H$, the scalar field oscillates with frequency equal to its mass. Also show that its energy density falls off as $a^{-3}$, so it behaves exactly like ordinary nonrelativistic matter. Use this to justify why we can ignore fields that are much heavier than $H$ during inflation.

**7.7** Derive some useful identities involving the slow-roll parameters during inflation.

**(a)** Show that

$$\frac{d}{d\eta} \left( \frac{1}{aH} \right) = \epsilon_{\rm sr} - 1.$$

**(b)** Show that

$$4\pi G (\bar{\phi}')^2 = \epsilon_{\rm sr} a^2 H^2. \tag{7.114}$$

**(c)** Using the definitions of $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$, show that

$$\frac{d\epsilon_{\rm sr}}{d\eta} = 2aH\epsilon_{\rm sr}(\epsilon_{\rm sr} + \delta_{\rm sr}). \tag{7.115}$$

Use this to show that

$$\left. \frac{d \ln \epsilon_{\rm sr}}{d \ln k} \right|_{aH=k} = 2(\epsilon_{\rm sr} + \delta_{\rm sr}). \tag{7.116}$$

**7.8** Express the slow-roll parameters $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$ in terms of the potential $V$ and its derivatives with respect to $\phi$. Show that, to lowest order,

$$\epsilon_{\rm sr} = \frac{1}{16\pi G}\left(\frac{V_{,\phi}}{V}\right)^2$$

and

$$\delta_{\rm sr} = \epsilon_{\rm sr} - \frac{1}{8\pi G}\frac{V_{,\phi\phi}}{V},$$

where the derivatives of the potential are evaluated at $\bar{\phi}$.

**7.9** There are a number of ways of describing pressure in the universe and of relating the pressure to the energy density. One was introduced back in Ch. 2, the *equation of state*,

$$w \equiv \frac{\mathcal{P}}{\rho}. \tag{7.117}$$

The second is the *sound speed* squared,

$$c_s^2 \equiv \frac{d\mathcal{P}}{d\rho}. \tag{7.118}$$

In the homogeneous universe, one computes $c_s^2$ by differentiating both $\mathcal{P}$ and $\rho$ with respect to time and taking the ratio: $c_s^2 = \dot{\mathcal{P}}/\dot{\rho}$. Finally, there is the ratio of perturbations in the pressure to those in the energy density,

$$\frac{\delta\mathcal{P}}{\delta\rho} = -\frac{\delta T^i{}_i}{3\delta T^0_0}, \tag{7.119}$$

where the minus sign accounts for the fact that the time-time component of the energy-momentum tensor is minus the energy density, and the factor of 3 negates the sum over the three spatial indices. For adiabatic perturbations, $\delta\mathcal{P}/\delta\rho = c_s^2$. Show that this holds for three separate cases: matter, radiation, and a single scalar field during inflation at the time of horizon crossing. For the last case, it is enough to show that the difference $\delta\mathcal{P}/\delta\rho - c_s^2$ is of order the slow-roll parameters $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$.

**7.10** Calculate some well-known properties of the quantized harmonic oscillator.

**(a)** The momentum of the harmonic oscillator with unit mass is $p = dx/dt$. Calculate

$$[\hat{x}, \hat{p}]$$

and show that it is equal to $i$. You can obtain the operator $\hat{p}$ by differentiating $\hat{x}$ (Eq. (7.20)) with respect to time.

**(b)** Calculate the zero-point energy of the harmonic oscillator with unit mass. Do this by quantizing the energy

$$E = \frac{p^2}{2} + \frac{\omega^2 x^2}{2}$$

and then computing its expectation value in the ground state: $\langle 0|\hat{E}|0\rangle$.

**7.11** Show that gravitational waves are not sourced by the scalar field during inflation. To do this, recall that the right-hand side of Eq. (7.27) is, assuming $h = h_+$,

$$\delta T^1{}_1 - \delta T^2{}_2$$

where $\delta T$ is the perturbation to the energy-momentum tensor (assumed to be dominated by $\phi$) and, as in the derivation of Eq. (6.73), we have chosen $\boldsymbol{k}$ to be in the $z$ direction. Show that this right-hand side is indeed zero for the scalar field.

**7.12** Show that Eq. (7.40) is the appropriate solution to Eq. (7.39).

**(a)** Define $\tilde{v} = v/\eta$ and rewrite Eq. (7.39) in terms of $\tilde{v}$.

**(b)** The resulting equation is the spherical Bessel equation. Write down the general solution to this as a linear combination of two functions of $k\eta$.

**(c)** Use the Minkowski-space solution for the harmonic oscillator for $k|\eta| \gg 1$ as initial condition to determine the coefficients of part (b). Show that this yields Eq. (7.40).

**7.13** Show that on large scales Eq. (7.66) holds. One way to do this is to combine Einstein's equations, the time-time (Eq. (6.41)) and time-space (Exercise 6.6) components, and to take the large-scale limit.

**7.14** Using the results of Sect. 7.6, derive the *consistency relation*—a robust prediction of single-field inflation—between the two observables $n_{\rm T}$ and $r$.

**7.15** Compute the ratio of neutrino to radiation energy density, $\mathfrak{f}_\nu \equiv \rho_\nu/\rho_\gamma$, after electron–positron annihilation (see Sect. 2.4.4) but before neutrino masses become relevant.

**7.16** Account for the neutrino quadrupole moment when setting up initial conditions during radiation domination, neglecting neutrino masses.

**(a)** Start with Eq. (5.73), and set $E_\nu(p) = p$. Then, $p$ disappears from the equation and we can set $\mathcal{N} = \mathcal{N}(k, \eta, \mu)$. Turn the equation into a hierarchy of equations for the neutrino moments $\mathcal{N}_l(k, \eta)$, truncated at $\mathcal{N}_2$:

$$\begin{aligned} \mathcal{N}_0' + k\mathcal{N}_1 &= -\Phi' \\ \mathcal{N}_1' - \frac{k}{3}(\mathcal{N}_0 - 2\mathcal{N}_2) &= \frac{k}{3}\Psi \\ \mathcal{N}_2' - \frac{2}{5}k\mathcal{N}_1 &= 0. \end{aligned} \tag{7.120}$$

To do this, you need to recall the definition of these moments, which is equivalent to that for photons, Eq. (5.66). A good way to reduce Eq. (5.73) into this hierarchy is to multiply it first by $\mathcal{P}_0$ and then integrate over $\mu$. This leads to the

first equation above. Then multiply Eq. (5.73) by $\mathcal{P}_1$ to get the second and $\mathcal{P}_2$ to get the third. More details are given in Sect. 9.3, where we go through the same exercise for the photon moments. In the third equation you may neglect $\mathcal{N}_3$ because it is smaller than $\mathcal{N}_2$ by a factor of order $k\eta$ (show this!).

**(b)** Eliminate $\mathcal{N}_1$ from these equations and show that

$$\mathcal{N}_2'' = \frac{2k^2}{15}(\Psi + \mathcal{N}_0 - 2\mathcal{N}_2). \tag{7.121}$$

Drop $\mathcal{N}_2$ on the right-hand side because it is much smaller than $\Psi + \mathcal{N}_0$.

**(c)** Rewrite Einstein's equation (6.48) as

$$\mathcal{N}_2 = -(k\eta)^2 \frac{\Phi + \Psi}{12\mathfrak{f}_\nu}, \tag{7.122}$$

where $\mathfrak{f}_\nu$ is defined in Exercise 7.15. This neglects the photon quadrupole, which is a reasonable assumption since Compton scattering sets $\Theta_2 \ll \mathcal{N}_2$.

**(d)** Now differentiate this form of Einstein's equation twice to get an expression for $\mathcal{N}_2''$. Equate this to the expression for $\mathcal{N}_2''$ derived in part (b). (You may drop all time derivatives of $\Phi$ and $\Psi$ when doing this since the mode of interest is the $p = 0$ constant mode.) Use this equation to express $\mathcal{N}_0$ in terms of $\Phi$ and $\Psi$.

**(e)** Finally assume that $\Theta_0 = \mathcal{N}_0$ and use your expression for $\mathcal{N}_0$ to rewrite Eq. (7.85) as a relation between the two gravitational potentials. Show that this relation is

$$\Phi = -\Psi\left(1 + \frac{2\mathfrak{f}_\nu}{5}\right). \tag{7.123}$$

**7.17** Show that the initial conditions for the velocities and dipoles of matter and radiation are as given in Eq. (7.95).

**7.18** Determine the predictions of an inflationary model with a quadratic potential:

$$V(\phi) = \frac{1}{2}m^2\phi^2. \tag{7.124}$$

**(a)** Compute the slow-roll parameters $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$ in terms of $\phi$. What can you already say about the tensor-to-scalar ratio $r$ and the spectral index $n_s$?

**(b)** Determine $\phi_e$, the value of the field at which inflation ends, by setting $\epsilon_{\rm sr} = 1$ at the end of inflation.

**(c)** To determine the spectrum of perturbations, you will need to evaluate $\epsilon_{\rm sr}$ and $\delta_{\rm sr}$ at $-k\eta = 1$, and hence $\phi$ at that epoch. Choose the wavenumber $k = k_{\rm p}$, and determine $\phi$ by relating the value of $aH$ when $k_{\rm p}$ left (and later re-entered) the horizon to $a_e H_e$, the value at the end of inflation. For this, assume $N$ $e$-folds from the epoch of horizon exit to the end of inflation. You can further assume that $H$ remains constant during inflation, and that the universe after inflation is radiation-dominated throughout.

**(d)** Match the values of $n_s$ and $\mathcal{A}_s$ that are predicted in this model to the values of our fiducial cosmology, and use this to determine $m/H_e$ and $N$. Finally, evaluate $m$, and $\phi$ at the epoch when $k_{\rm p}$ leaves the horizon, in units of the Planck mass. What is the predicted amplitude of tensor modes in this case? Compare with Fig. 10.11.

This model illustrates the features of *large-field* inflationary models: the field value is of order or even greater than $m_{\rm Pl}$, but the energy scale $V$ is much smaller than $m_{\rm Pl}^4$.

# 8 Growth of structure: linear theory

Having set up the system of equations to be solved and the initial conditions for the perturbations, we can now compute the inhomogeneities in the matter and anisotropies in the photons. In this chapter, we focus on the perturbations to the dark matter: the density perturbation $\delta_c$ and velocity $u_c$. These are coupled to all other perturbations only via gravity. For this reason, perturbations to the dark matter depend very little on the details of the radiation perturbations: at late times, when the universe is dominated by matter, the potentials $\Phi$, $\Psi$ which mediate the effect of gravity are independent of the radiation. At early times, the dominant radiation perturbations are relatively simple, so that all moments beyond the monopole and dipole can be neglected. Our rationale for starting with matter is that the converse is not true, as we will see in the next chapter: to treat the anisotropies in the radiation properly, we will need to know how the matter perturbations behave.

The ultimate goal of this chapter is to obtain a prediction for the linear matter power spectrum. We will obtain this by solving for the evolution of each Fourier mode. Given this solution, and the initial power spectrum generated by inflation, we can construct the power spectrum of matter as a function of redshift. On large scales, this can be compared with observations of galaxy clustering and lensing, as we will see in Ch. 11 and Ch. 13. Matter becomes nonlinear on small scales in the late universe however, so our results cannot be directly compared with observations on those scales. However, they serve as starting point for analytic and numerical calculations of nonlinear structure, which are the topic of Ch. 12.

Several publicly available codes[1] calculate the transfer function to subpercent precision in a matter of seconds by numerically solving the equations we derived in previous chapters. In Sect. 8.2 and Sect. 8.3 of this chapter, we develop approximate analytic solutions. They yield additional physical insight into the physics of structure growth in the early universe, but going through these sections is not required in order to follow the rest of the book. The main features of the calculation are summarized in Sect. 8.1.

## 8.1 Prelude

Gravitational instability is fundamentally responsible for the structure in our universe. As time evolves, matter accumulates in initially slightly overdense regions. Despite the fact that initial overdensities were of very small amplitude (of order $10^{-4}$), eventually enough matter accumulated over the age of the universe to form the very significant structures we see in the universe today.

[1] The most popular ones are *CAMB* (Lewis et al., 2000) and *CLASS* (Blas et al., 2011).

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00014-0
Copyright © 2021 Elsevier Inc. All rights reserved.

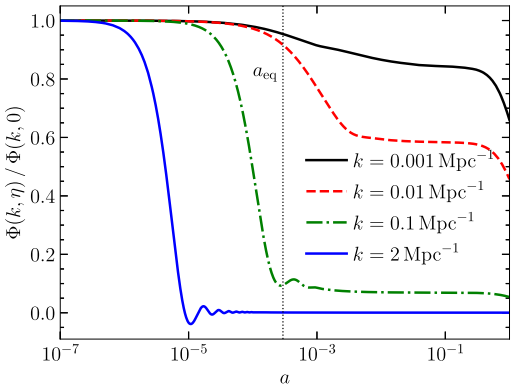



**FIGURE 8.1** The linear evolution of the gravitational potential $\Phi$ for modes of different wavenumber in the fiducial $\Lambda$CDM cosmology. In each case, we have normalized to the value of the potential at early times.

Apart from the attractive force of gravity, there are two counteracting effects: first, the expansion of the background universe, which tends to drag particles of all species apart. The faster the expansion, the slower the growth of structure. In non-expanding space, a small overdensity will grow exponentially fast under gravity (if there are no pressure perturbations; see below); in the expanding universe, this exponential growth is slowed down to a power-law or even logarithmic growth in time. In particular, we will see that structure grows more slowly during radiation domination than later during matter domination and finally slows down again once dark energy begins to dominate.

The second effect is specific to baryons and photons, which exert pressure: pressure increases in proportion to density, and gas tends to move in the direction of lower pressure (opposite to the pressure gradient). This means that an overdensity in the baryons does not accumulate matter as quickly as one in the dark matter, since the larger pressure compared to the environment tends to slow down or stop inflowing gas.

In this chapter we will treat super-horizon ($k\eta \ll 1$) versions of gravitational growth as well as the more familiar sub-horizon version ($k\eta \gg 1$), both with and without perturbations in the radiation component. While going through the math, it is useful to bear in mind the dueling forces of gravity, expansion, and pressure perturbations.

### 8.1.1 Three stages of evolution

The evolution of cosmological perturbations breaks up naturally into three stages. To see this, let us cheat and look at the solutions for several different modes. Fig. 8.1 shows the gravitational potential as a function of scale factor for long-, medium-, and short-wavelength modes. Early on, all of the modes are outside the horizon ($k\eta \ll 1$; recall that $\eta$ is positive after inflation) and the potential is constant. At intermediate times, two things happen: the wavelengths enter the horizon and the universe evolves from radiation domination ($a \ll a_{\rm eq}$) to matter domination ($a \gg a_{\rm eq}$). Fig. 8.1 illustrates that the order of these epochs ($a_{\rm eq}$ and the epoch of horizon crossing) greatly affects the potential. The large-scale

mode, which enters the horizon well after $a_{\rm eq}$, evolves much differently from small-scale modes, which enter the horizon before equality. Finally, at late times, all the modes evolve identically again, remaining constant during matter domination before decaying once dark energy becomes important.

We are able to observe the distribution of matter predominantly at late epochs, in the third stage of evolution, when all modes are evolving identically. If we wish to relate the potential during these times to the primordial curvature perturbation $\mathcal{R}$ generated during inflation, we can write schematically

$$\Phi(\boldsymbol{k}, a) = \frac{3}{5}\mathcal{R}(\boldsymbol{k}) \times \left\{\text{Transfer Function}\,(k)\right\} \times \left\{\text{Growth factor}\,(a)\right\}. \tag{8.1}$$

We will get to understand the $3/5$ factor in a little while. The transfer function describes the evolution of perturbations through the epochs of horizon crossing and radiation/matter transition, while the growth factor describes the wavelength-independent growth at late times. This schematic equation is indeed how the growth factor and the transfer function are defined, with two bows to convention. Notice from Fig. 8.1 that even the largest-wavelength perturbations decline slightly as the universe passes through the epoch of equality. This decline is conventionally removed so that the transfer function on large scales is equal to 1. Therefore, the transfer function is defined as

$$T(k) \equiv \frac{\Phi(\boldsymbol{k}, a_{\rm late})}{\Phi_{\text{large-scale}}(\boldsymbol{k}, a_{\rm late})} \tag{8.2}$$

where $a_{\rm late}$ denotes an epoch deep in matter domination, and the *large-scale* solution is the primordial $\Phi$ decreased by a small amount; strictly, it is the solution of the gravitational potential for modes that entered the horizon well in the matter-dominated epoch. We will derive in Sect. 8.2 that—neglecting anisotropic stress—this factor is equal to $9/10$. The second convention concerns the growth factor. The ratio of the potential to its value right after the transfer function regime is defined to be

$$\frac{\Phi(\boldsymbol{k}, a)}{\Phi(\boldsymbol{k}, a_{\rm late})} \equiv \frac{D_+(a)}{a} \qquad (a > a_{\rm late}), \tag{8.3}$$

where $D_+$ is called the growth factor. During matter domination, the potential is constant so $D_+(a) = a$. With these conventions, we have

$$\Phi(\boldsymbol{k}, a) = \frac{3}{5}\mathcal{R}(\boldsymbol{k})T(k)\frac{D_+(a)}{a} \qquad (a > a_{\rm late}). \tag{8.4}$$

The evolution of the CDM overdensity of matter follows from the evolution of $\Phi$, as depicted in Fig. 8.2 for four different modes. Notice that at late times—when the potential is constant and all the modes are within the horizon—the overdensity grows in time: $\delta_{\rm c}(\boldsymbol{k}, a) \propto D_+(a)$. This explains the seemingly odd nomenclature above (why is it called a *growth* factor if the potential remains constant?): $D_+$ describes the growth of the matter

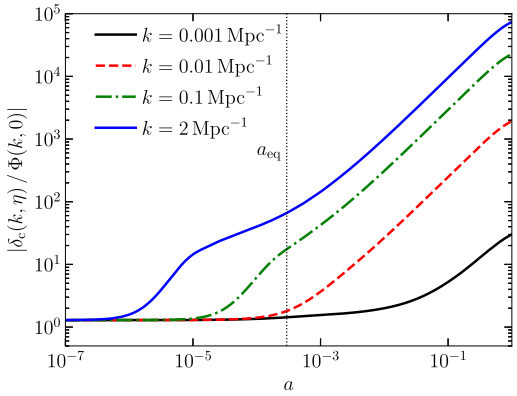



**FIGURE 8.2** The evolution of dark matter density perturbations in the fiducial ΛCDM cosmology. We have normalized to the potential at early times as in Fig. 8.1. The amplitude of each mode starts to grow upon horizon entry. Well after $a_{\rm eq}$, all sub-horizon modes evolve identically, and scale as the growth factor $D_+(a)$. During matter domination, before Λ becomes relevant, $D_+(a) = a$. At the very latest times, we can see a slight suppression from this linear trend due to the onset of accelerated expansion.

perturbations at late times. This growth is completely consistent with our intuition that, as time evolves, overdense regions attract more and more matter, thereby becoming more overdense.

In the late universe, baryons closely follow the dark matter, so we typically describe them together in form of the total matter overdensity $\delta_{\rm m}$. So let us now express the power spectrum of the matter distribution in terms of the primordial power spectrum generated during inflation, the transfer function, and the growth factor. The simplest way to relate the matter overdensity to the potential at late times is to use Poisson's equation (6.80) in the large-$k$, no-radiation limit,

$$k^2\Phi(\boldsymbol{k},a) = 4\pi G\rho_{\rm m}(a)a^2\delta_{\rm m}(\boldsymbol{k},a) \qquad (a > a_{\rm late},\ k \gg aH)\,. \tag{8.5}$$

This equation is no longer correct if $k$ is of order $aH$ or less. For large-scale structure applications, this is not a big worry, as the most precise measurements are for modes that satisfy $k \gg aH$.[2]

Now, the background density of matter (including baryons) is $\rho_{\rm m} = \Omega_{\rm m}\rho_{\rm cr}/a^3$, and $4\pi G\rho_{\rm cr} = (3/2)H_0^2$, so

$$\delta_{\rm m}(\boldsymbol{k},a) = \frac{2k^2a}{3\Omega_{\rm m}H_0^2}\Phi(\boldsymbol{k},a) \qquad (a > a_{\rm late},\ k \gg aH). \tag{8.6}$$

[2] Moreover, Eq. (8.5) *does* hold on all scales if $\delta_{\rm m}$ on the right-hand side is defined in synchronous-comoving gauge (see Exercise 5.1). The density in this gauge is in many cases more directly related to observables and simulations than $\delta_{\rm m}$ in conformal-Newtonian gauge.

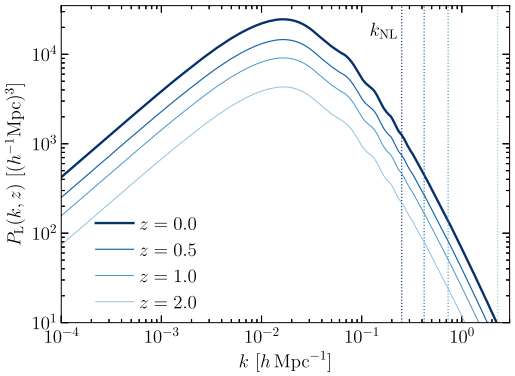



**FIGURE 8.3** The linear matter power spectrum in the fiducial ΛCDM cosmology at different redshifts. Scales to the left of the vertical lines, which indicate $k_{\rm NL}(z)$ for each of the redshifts shown, are still evolving approximately linearly at each redshift.

This, together with Eq. (8.4), allows us to relate the overdensity in the late universe to the primordial potential:

$$\delta_{\rm m}(\boldsymbol{k},a) = \frac{2}{5}\frac{k^2}{\Omega_{\rm m}H_0^2}\mathcal{R}(\boldsymbol{k})T(k)D_+(a) \qquad (a > a_{\rm late},\ k \gg aH). \tag{8.7}$$

Eq. (8.7) holds regardless of how the initial perturbation $\mathcal{R}$ was generated, as long as it is an adiabatic perturbation. In the context of inflation, we saw in the previous chapter that $\mathcal{R}(\boldsymbol{k})$ is drawn from a Gaussian distribution with mean zero and power spectrum $P_{\mathcal{R}}(k) = (2\pi^2/k^3)\mathcal{A}_s(k/k_{\rm p})^{n_s-1}$ (Eq. (7.99)). So the linear power spectrum of matter at late times is

$$P_{\rm L}(k,a) = \frac{8\pi^2}{25}\frac{\mathcal{A}_s}{\Omega_{\rm m}^2}D_+^2(a)T^2(k)\frac{k^{n_s}}{H_0^4 k_{\rm p}^{n_s-1}}. \tag{8.8}$$

Notice that $(i)$ the power spectrum has dimensions of (length)$^3$; and $(ii)$ Eq. (8.8) implies that $P_{\rm L}(k) \propto k^{n_s}$ on large scales where $T(k) = 1$.

Fig. 8.3 shows the matter power spectrum for our fiducial ΛCDM cosmology, today as well as at higher redshifts. While on large scales we see the expected behavior, on small scales the power spectrum turns over. To understand this, look back at Fig. 8.1. The small-scale mode there ($k = 2\,h$ Mpc$^{-1}$) enters the horizon well before matter/radiation equality. During the radiation epoch the potential decays, so the transfer function is much smaller than unity. The effect of this on matter perturbations can be seen in Fig. 8.2, where the growth of $\delta$ is retarded starting at $a \simeq 10^{-5}$ after the mode has entered the horizon and ending at $a \simeq 10^{-4}$ when the universe becomes matter dominated. Modes that enter the horizon even earlier undergo more suppression. Thus, the power spectrum is a decreasing function of $k$ on small scales. This leads to the realization that there will be a turnover in the power spectrum at a scale $k_{\rm eq}$ corresponding to the one which enters the horizon at

matter/radiation equality. Measuring this scale thus allows us to constrain the amount of matter in the universe.

Another important scale to keep in mind is the scale $k_{\rm NL}$ above which nonlinearities cannot be ignored. To estimate this, we use the variance of (linear) density perturbations $\Delta^2_{\rm L}(k)$ generated by modes within a logarithmic wavenumber bin $d\ln k$ centered around $k$. We have

$$\Delta^2_{\rm L}(k,a) = \frac{1}{\epsilon}\int_{|\ln k'-\ln k|<\epsilon}\frac{d^3k'}{(2\pi)^3}P_{\rm L}(k',a) = \int_{|\ln k'-\ln k|<\epsilon}k'^3\frac{dk'}{k'}\int\frac{d\Omega'}{(2\pi)^3}P_{\rm L}(k',a)$$
$$= \frac{k^3P_{\rm L}(k,a)}{2\pi^2}, \tag{8.9}$$

since the integral over $d\Omega'$ simply gives $4\pi$. In the second line we have used the fact that we are considering an infinitesimal wavenumber bin. A regime where $\Delta^2_{\rm L}\ll 1$ then corresponds to small inhomogeneities, while $\Delta^2_{\rm L}\gtrsim 1$ indicates nonlinear perturbations. When plotted vs. $\ln k$, $\Delta^2_{\rm L}(k)$ immediately tells us on what scales we expect significant density fluctuations (Fig. 12.1 shows an analogous plot of the variance of the density field). Solving the condition $\Delta^2_{\rm L}(k_{\rm NL},a)\simeq 1$ for $k_{\rm NL}$ yields $k_{\rm NL}(a=1)\simeq 0.25\,h\,{\rm Mpc}^{-1}$ today. At earlier times, structure was not as evolved, so the nonlinear scale was smaller, or equivalently $k_{\rm NL}(a)$ was larger in the past (see the dashed vertical lines in Fig. 8.3). The power spectrum shown in Fig. 8.3 is the linear power spectrum. On scales approaching $k_{\rm NL}$, this is just a hypothetical quantity, and one cannot directly compare $P_{\rm L}(k,a)$ with the matter distribution. We will return to this issue in Ch. 12.

### 8.1.2 Closing the Boltzmann hierarchy

What are the evolution equations for the dark matter overdensity? Since all constituents are coupled by gravity, in principle these are the full set of Boltzmann equations derived in Ch. 5 and the pair of Einstein equations from Ch. 6. To get a qualitative understanding though, the full set of equations is not needed. To understand why, recall that early on (before recombination at $a=a_*$), the photon distribution can be characterized by only two moments, the monopole $\Theta_0$ and the dipole $\Theta_1$. All other moments are suppressed because the photons are tightly coupled to the electron/proton plasma. After decoupling this ceases to be true, and to completely characterize the photon distribution we will need to follow higher-order moments. However, for the purposes of the matter distribution, what the photons are doing after $a_*$ is irrelevant. For, by that time, which is well into the matter era, the potential is dominated by the dark matter itself. To sum up then, we can neglect all photon moments except for the monopole and dipole when we are considering the evolution of the matter distribution.

In the following, we will also neglect the higher multipoles for neutrinos, since they are more difficult to treat analytically. But, as neutrinos free-stream instead of being tightly coupled, this is also inaccurate. Nevertheless, neglecting the higher neutrino moments is still more accurate than neglecting neutrinos entirely. Hence, we proceed with

the monopole $\Theta_{\mathrm{r},0}$ and dipole $\Theta_{\mathrm{r},1}$ of the *total* radiation, defined in Eq. (6.79), not to be confused with those of the photon distribution $\Theta_0$, $\Theta_1$. When neglecting the higher radiation moments, the Boltzmann equations for neutrinos and photons become identical; moreover, the photon and neutrino distributions start with the same adiabatic initial conditions, so we can combine them. Supplementing these with the equations for the dark matter, we obtain our desired set of equations, based on Sect. 5.7:

$$\Theta_{\mathrm{r},0}' + k\Theta_{\mathrm{r},1} = -\Phi', \tag{8.10}$$

$$\Theta_{\mathrm{r},1}' - \frac{k}{3}\Theta_{\mathrm{r},0} = -\frac{k}{3}\Phi, \tag{8.11}$$

$$\delta_{\mathrm{c}}{}' + iku_{\mathrm{c}} = -3\Phi', \tag{8.12}$$

$$u_{\mathrm{c}}{}' + \frac{a'}{a}u_{\mathrm{c}} = ik\Phi. \tag{8.13}$$

Even with the assumption that only the monopole and dipole are retained, getting from Eq. (5.67) to Eqs. (8.10) and (8.11) requires some work. In particular, we have used the tight-coupling approximation, which allows us to eliminate the baryon perturbations. This is a fairly good approximation since the baryons are kept close to uniform at early times via their tight coupling to photons (we will explore the effects of baryons in Sect. 8.6). In Exercise 8.1, you can work out the steps in detail. Note that these simplifications are employed merely because we are only interested in the CDM perturbations in this chapter; we will explore the full photon evolution equation in the next chapter.

To close the set of equations for the dark matter density, we need an equation for the gravitational potential $\Phi$. You may have noticed that in Eq. (8.10) and following we set $\Psi \to -\Phi$, an approximation valid in the limit that there are no quadrupole moments (Eq. (6.48)). Since some of the Einstein equations are redundant, we have several choices for an equation relating $\Phi$ to the radiation and matter overdensities. We can use the time-time component, Eq. (6.41),

$$k^2\Phi + 3\frac{a'}{a}\left(\Phi' + \frac{a'}{a}\Phi\right) = 4\pi G a^2[\rho_{\mathrm{c}}\delta_{\mathrm{c}} + 4\rho_{\mathrm{r}}\Theta_{\mathrm{r},0}]. \tag{8.14}$$

Here, we have again set $\Psi$ to $-\Phi$, neglected the baryons following the arguments above, and merged the neutrino and photon contributions to the potential. The alternative is to use the algebraic (no time derivatives) equation (6.80):

$$k^2\Phi = 4\pi G a^2\left[\rho_{\mathrm{c}}\delta_{\mathrm{c}} + 4\rho_{\mathrm{r}}\Theta_{\mathrm{r},0} + \frac{3aH}{k}\left(i\rho_{\mathrm{c}}u_{\mathrm{c}} + 4\rho_{\mathrm{r}}\Theta_{\mathrm{r},1}\right)\right]. \tag{8.15}$$

Both of these equations will be useful to us at various times, although only one is necessary to close the set of equations for the five variables $\delta_{\mathrm{c}}$, $u_{\mathrm{c}}$, $\Theta_{\mathrm{r},0}$, $\Theta_{\mathrm{r},1}$, and $\Phi$.

At this stage, the most straightforward way to proceed is to solve the set of five coupled equations numerically (Exercise 8.2). If Eq. (8.14) is used, there are no numerical difficulties, and with very little work, you can have a code that computes the transfer function (in the absence of baryons) in less than a second.

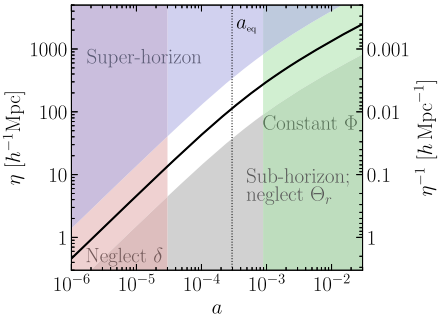



**FIGURE 8.4** Physics regimes of the transfer function. Shaded regions show different regimes where analytic expressions exist that we will derive in Sects. 8.2–8.3. The gap in the center shows that no analytic solutions exist to capture the full evolution of intermediate-scale modes. The curve monotonically increasing from bottom left to top right is the comoving horizon $\eta(a)$, with axis on the left; the right axis shows the corresponding wavenumber that crosses the horizon, i.e. $k = 1/\eta$.

Analytical solutions for the dark matter density are harder to come by. There is no analytic solution valid on all scales at all times. To make progress, we will have to take some limits that reduce the full set of five equations to a more manageable two or three. The cost is that these limits will be valid only for certain scales at certain times, so we have to patch these analytic solutions together to obtain a reasonable transfer function for all $k$. As we mentioned, extremely accurate codes that calculate the transfer function in a matter of seconds are freely available and easy to use. Rather than obtaining precise numerical results, our goal for this chapter is thus to develop an understanding of the relevant physics in different regimes, and to see how far this takes us in reproducing the numerical results.

As a guide to this analytic work which will occupy us in Sect. 8.2–8.3, consider Fig. 8.4. The solid curve is the comoving horizon (i.e. the conformal time $\eta$), which increases with time, equal to about $110\,h^{-1}$ Mpc at the epoch of equality in our fiducial cosmology. A given comoving scale remains constant with time. There are several regimes where we can make progress by employing physical approximations:

- The ***super-horizon regime***, when $k\eta \ll 1$, allows for exact solutions for the entire time evolution (Sect. 8.2.1).
- The ***horizon entry***, when $\eta$ has increased sufficiently so that $k\eta > 1$, can be described if it happens at late times, i.e. for large-scale modes (in which case $\Phi$ remains constant, Sect. 8.2.2), or at very early times during radiation domination, which applies to small-scale modes (neglecting $\delta$, Sect. 8.3.1).
- The ***sub-horizon evolution*** of large-scale modes is covered by the constant-$\Phi$ solution of Sect. 8.2.2. For small-scale modes, we will find another approximation, neglecting radiation perturbations, in Sect. 8.3.2.

None of these approximate solutions is able to describe modes that enter the horizon around matter-radiation equality (the white spot in the center of Fig. 8.4). This is merely due to a lack of simplifying limits we can employ; the physics describing the evolution of

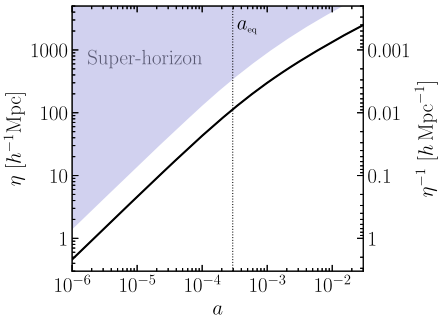



**FIGURE 8.5** The regime studied in Sect. 8.2.1: super-horizon evolution of perturbations which remain outside the horizon until matter domination.

these modes is the same as that of smaller- and larger-scale modes. Indeed, the final transfer function we present in Sect. 8.4 is a smooth function of scale.

## 8.2 Large scales

On large scales, we can get analytic solutions for the potential first through the matter-radiation transition and then through horizon crossing. We start with the super-horizon solution valid through the matter-radiation transition. The result of Sect. 8.2.1 will be that the potential drops by a factor of 9/10 as the universe goes from radiation to matter domination.

### 8.2.1 Super-horizon solution

For modes that are far outside the horizon, $k\eta \ll 1$, the regime highlighted in Fig. 8.5, we can drop all terms in the evolution equations that depend on $k$. From Eq. (8.10) and Eq. (8.12), we see that, in this limit, the velocities ($u_{\rm c}$ and $\Theta_{\rm r,1}$) decouple from the evolution equations. This immediately reduces the number of equations to solve from five to three. For the third equation, we notice that Eq. (8.15) has terms inversely proportional to $k$. These will be difficult to deal with, so let us choose Eq. (8.14) instead. We are left with

$$\Theta'_{\rm r,0} = -\Phi', \tag{8.16}$$

$$\delta_{\rm c}{}' = -3\Phi', \tag{8.17}$$

$$3\frac{a'}{a}\left(\Phi' + \frac{a'}{a}\Phi\right) = 4\pi G a^2[\rho_{\rm c}\delta_{\rm c} + 4\rho_{\rm r}\Theta_{\rm r,0}]. \tag{8.18}$$

We can go a step further by realizing that the first two equations require $\delta_{\rm c} - 3\Theta_{\rm r,0}$ to be constant. Further, we know that this constant is zero (these are the adiabatic initial conditions from Sect. 7.5). So let us use the dark matter equation (8.17) and the Einstein equation

with $\Theta_{r,0}$ set to $\delta_c/3$. The Einstein equation is then

$$3\frac{a'}{a}\left(\Phi' + \frac{a'}{a}\Phi\right) = 4\pi G a^2 \rho_c \delta_c \left[1 + \frac{4}{3y}\right], \tag{8.19}$$

where we have introduced

$$y \equiv \frac{a}{a_{eq}} = \frac{\rho_m}{\rho_r}, \tag{8.20}$$

which we will use as an evolution variable instead of $\eta$ or $a$. Since we are ignoring baryons, we could also replace the numerator in Eq. (8.20) with $\rho_c$, a simple fix to slightly improve the accuracy of our analytical solutions.

Eqs. (8.17) and (8.19) are two first-order equations for the two variables $\delta_c$ and $\Phi$. The strategy will be to turn these two first-order equations into one second-order equation and then solve. First, though, let us rewrite the equations in terms of the new variable $y$. The derivative with respect to $y$ is related to that with respect to $\eta$ via

$$\begin{aligned}\frac{d}{d\eta} &= \frac{dy}{d\eta}\frac{d}{dy} \\ &= aHy\frac{d}{dy},\end{aligned} \tag{8.21}$$

where the second line follows from the definition of $y$ and the fact that $a' = a^2 H$. In terms of $y$ then, the Einstein equation becomes

$$\begin{aligned}y\frac{d\Phi}{dy} + \Phi &= \frac{y}{2(y+1)}\delta_c\left[1 + \frac{4}{3y}\right] \\ &= \frac{3y+4}{6(y+1)}\delta_c\end{aligned} \tag{8.22}$$

where the right-hand side of the first line follows since $8\pi G\rho_c/3 = (8\pi G\rho/3)y/(y+1) = H^2 y/(y+1)$.

In general, to turn two first-order equations into one second-order equation, we differentiate one of them. Here, to simplify the algebra, we first rewrite Eq. (8.22) as an expression for $\delta_c$; then differentiate with respect to $y$; and finally set $d\delta_c/dy$ to $-3d\Phi/dy$ thanks to the dark matter equation (8.17). This leads to

$$-3\frac{d\Phi}{dy} = \frac{d}{dy}\left\{\frac{6(y+1)}{3y+4}\left[y\frac{d\Phi}{dy} + \Phi\right]\right\}. \tag{8.23}$$

Carrying out the derivative is tedious but straightforward. We are left with

$$\frac{d^2\Phi}{dy^2} + \frac{21y^2 + 54y + 32}{2y(y+1)(3y+4)}\frac{d\Phi}{dy} + \frac{\Phi}{y(y+1)(3y+4)} = 0. \tag{8.24}$$

Remarkably, Kodama and Sasaki (1984) found an analytic solution to Eq. (8.24). They introduced a new variable

$$u \equiv \frac{y^3}{\sqrt{1+y}}\Phi. \tag{8.25}$$

In terms of this variable, you will show (Exercise 8.4) that Eq. (8.24) becomes

$$\frac{d^2u}{dy^2} + \frac{du}{dy}\left[-\frac{2}{y} + \frac{3/2}{1+y} - \frac{3}{3y+4}\right] = 0. \tag{8.26}$$

That is, there is no term proportional to $u$. Instead of a second-order equation for $\Phi$, then, we have a first-order equation for $du/dy$. Fortunately, this first-order equation is integrable. Denoting $u' \equiv du/dy$ for the next few steps, we have

$$\frac{du'}{u'} = dy\left[\frac{2}{y} - \frac{3/2}{1+y} + \frac{3}{3y+4}\right], \tag{8.27}$$

which we can integrate to get

$$\ln(u') = 2\ln(y) - (3/2)\ln(1+y) + \ln(3y+4) + \text{constant}. \tag{8.28}$$

Then, exponentiating gives

$$u' = \frac{du}{dy} = A\frac{y^2(3y+4)}{(1+y)^{3/2}} \tag{8.29}$$

where $A$ is a constant to be determined.

We are one integral away from an analytic expression for the gravitational potential. Remembering the definition of $u$ (Eq. (8.25)), we can integrate Eq. (8.29) to obtain

$$\frac{y^3}{\sqrt{1+y}}\Phi = A\int_0^y d\tilde{y}\frac{\tilde{y}^2(3\tilde{y}+4)}{(1+\tilde{y})^{3/2}}. \tag{8.30}$$

Note that there should be another constant, $u(0)$, here. However, since $y^3\Phi \to 0$ early on, this constant vanishes. By similar logic, we can determine the constant $A$ even before performing the integral. For small $y$, the integrand becomes $4\tilde{y}^2$, so for small $y$, Eq. (8.30) becomes $\Phi = 4A/3$. Therefore, $A = 3\Phi(0)/4$. The integral can be done analytically (Exercise 8.4 again) leaving

$$\Phi(\boldsymbol{k}, y) = \frac{1}{10y^3}\left[16\sqrt{1+y} + 9y^3 + 2y^2 - 8y - 16\right]\Phi(\boldsymbol{k}, 0). \tag{8.31}$$

Eq. (8.31) is our final expression for the potential on super-horizon scales. Although it is not obvious, at small $y$ this expression sets $\Phi = \Phi(0)$, a constant. This must be so, since we chose the two constants of integration with precisely this condition. At large $y$, once the universe has become matter-dominated, the $y^3$ term in the brackets dominates, so

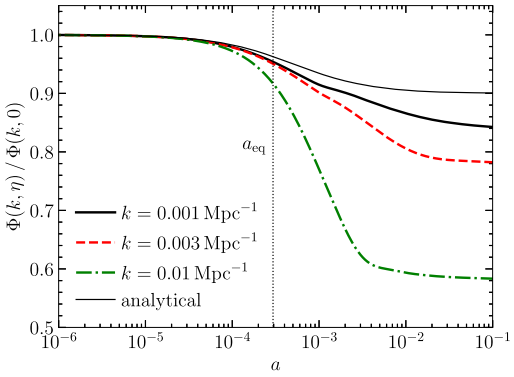



**FIGURE 8.6** Evolution of the potential in the fiducial ΛCDM cosmology, focusing on large-scale modes that are comparable to or larger than the horizon up until recombination. The thin solid line shows the analytic result of Eq. (8.31) which is valid only for modes much larger than the horizon (and neglects the neutrino and photon moments with $l \geq 2$).

$\Phi \to (9/10)\Phi(0)$. So, the potential on even the largest scales drops by $9/10$ as the universe passes through the epoch of equality.

This result allows us to obtain some very useful relations between the super-horizon gravitational potential $\Phi$ and the curvature perturbation $\mathcal{R}$. In Sect. 7.5, we derived $\Phi \simeq -\Psi = (2/3)\mathcal{R}$ during radiation domination. Now we have seen that $\Phi$ drops by a factor $9/10$ deep in matter domination. Since $\mathcal{R}$ is always conserved outside the horizon, we have $\Phi = (9/10)(2/3)\mathcal{R}$ in matter domination. So, to summarize

$$\Phi(\boldsymbol{k},\eta)\Big|_{\text{super-horizon}} = \begin{cases} \frac{2}{3}\mathcal{R}(\boldsymbol{k}), & \text{radiation domination,} \\ \frac{3}{5}\mathcal{R}(\boldsymbol{k}), & \text{matter domination.} \end{cases} \tag{8.32}$$

During matter domination, $\Phi$ does not even evolve inside the horizon as we mentioned, so that the relation $\Phi = (3/5)\mathcal{R}$ remains valid inside the horizon. This explains the $3/5$ factor we have included in the transfer function definition, Eq. (8.4).

Let us now compare our analytic result, valid only when modes are super-horizon, with the numerical results. Fig. 8.6 shows that the solution works reasonably well on the largest scales, the deviations being mostly due to the neutrino quadrupole $\mathcal{N}_2$ which we have neglected in our analytic calculation ($\mathcal{N}_2$ also leads to $\Phi$ being not exactly equal and opposite to $\Psi$). A feature of the analytic solution that may be surprising to you is that, although it is true that the large-scale potentials are constant in both the matter and radiation epochs, the transition between the pure matter and pure radiation eras is quite long. For example, and this is important for the purposes of the CMB as we will see in the next chapter, the potentials, even for the largest-scale modes, are still decaying as late as $a \simeq 10^{-2}$, significantly after $a_{\rm eq}$.

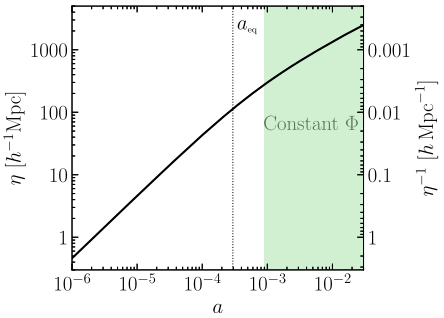



**FIGURE 8.7** The regime studied in Sect. 8.2.2: evolution of modes through horizon entry during matter domination.

## 8.2.2 Through horizon crossing

One interesting feature of Fig. 8.6 is that the large-scale potential (the numerical solution) becomes constant at very late times ($a \gtrsim 10^{-2}$). For $k = 10^{-3}\,h$ Mpc$^{-1}$, the mode enters the horizon at $\eta \sim k^{-1} = 1000\,h^{-1}$ Mpc which corresponds to $a \sim 0.006$ in our fiducial cosmology. This is the regime highlighted in Fig. 8.7. The potential remains constant as the mode crosses the horizon. This result is valid as long as the universe is matter dominated. We now set out to prove this.

We are interested then in our set of five equations in the limit that radiation is not important. The potential depends only on the matter inhomogeneities, so we can neglect the two radiation equations, (8.10) and (8.11). In addition to the two matter equations, we now keep the second of Einstein's equations (8.15). This is an algebraic equation, meaning that we could in principle eliminate $\Phi$ in the two matter equations and be left with a system of two first-order differential equations. These two first-order equations in general have two solutions. Instead of solving them directly, though, we can cheat using our knowledge of the initial conditions. Here is the idea: we just learned that, deep in the matter epoch, super-horizon potentials are constant. Therefore, the initial conditions for our problem are that the potential is constant ($\Phi' = 0$). If we can show that constant $\Phi$ is one of the two general solutions to the set of matter-dominated equations, then we do not care what the other solution is (we would have to care if this solution was growing, $\Phi' > 0$; however, it is in fact decaying). The initial conditions then ensure that the constant $\Phi$ solution will be *the* solution.

We want to see, then, if the set of equations

$$\delta_{\rm c}{}' + iku_{\rm c} = 0, \tag{8.33}$$

$$u_{\rm c}{}' + aHu_{\rm c} = ik\Phi, \tag{8.34}$$

$$k^2\Phi = \frac{3}{2}a^2H^2\left[\delta_{\rm c} + \frac{3aHiu_{\rm c}}{k}\right], \tag{8.35}$$

admits a solution with $\Phi$ being a constant in time. We can use the algebraic equation (8.35) to eliminate $\delta_{\rm c}$ from the other two equations. In the matter-dominated era, $H \propto a^{-3/2}$, so

$d(aH)/d\eta = -a^2H^2/2$. Replacing $\delta_c$ in Eq. (8.33) with $\Phi$ and $u_c$ therefore leads to

$$\frac{2k^2\Phi'}{3a^2H^2} + \frac{2k^2\Phi}{3aH} - \frac{3aHiu_c'}{k} + \frac{3a^2H^2iu_c}{2k} + iku_c = 0. \tag{8.36}$$

We now have two first-order equations for $\Phi$ and $u_c$. The strategy is to turn these two equations into one second-order equation for $\Phi$. First eliminate $u_c'$ from Eq. (8.36) by using the velocity equation. This leaves

$$\frac{2k^2\Phi'}{3a^2H^2} + \left[\frac{iu_c}{k} + \frac{2\Phi}{3aH}\right]\left(\frac{9a^2H^2}{2} + k^2\right) = 0. \tag{8.37}$$

If the second-order equation is of the form $\alpha\Phi'' + \beta\Phi' = 0$, that is, if it has no terms proportional to $\Phi$, then $\Phi =$ constant is a solution to the equations. So we differentiate Eq. (8.37) with respect to $\eta$ but consider only the terms proportional to $\Phi$, dropping all terms proportional to derivatives of $\Phi$. Using the fact that $(d/d\eta)(aH)^{-1} = 1/2$ during matter domination, we see that the remaining terms are

$$\left[\frac{iu_c'}{k} + \frac{\Phi}{3}\right]\left(\frac{9a^2H^2}{2} + k^2\right) + \left[\frac{iu_c}{k} + \frac{2\Phi}{3aH}\right]\frac{d}{d\eta}\frac{9a^2H^2}{2}$$
$$= -\left[\frac{iaHu_c}{k} + \frac{2\Phi}{3}\right]\left(9a^2H^2 + k^2\right) \tag{8.38}$$

where we have eliminated $u_c'$ by using the velocity equation again. But Eq. (8.37) tells us that the term in square brackets on the right here is proportional to $\Phi'$. So there are no terms proportional to $\Phi$. Constant potentials are therefore a solution in the matter-dominated era. Since the initial conditions pick out this mode, $\Phi =$ const is *the* solution in the matter-dominated era.

Potentials remain constant as long as the universe is matter dominated. This answers a question that arises in the context of an expanding universe: do potential wells grow deeper as more and more matter accretes into overdense regions? Or do they decay as matter is pulled apart by the expansion of the universe? The verdict is that, in a matter-dominated universe, the two effects delicately balance one another and potentials remain constant. When dark energy comes to dominate ($a \gtrsim 0.1$), this balance is destroyed and the potentials will decay. This decay is accurately described by the growth factor (Sect. 8.5), and does not affect the transfer function by construction. The main result of this section is that the transfer function as defined in Eq. (8.2) is very close to unity on all scales that enter the horizon after the universe becomes matter dominated. That is, it is unity for all $k \ll a_{\rm eq}H(a_{\rm eq})$, the inverse comoving Hubble radius at equality. You will show in Exercise 8.5 that the relevant scale is

$$k_{\rm eq} = 0.073\ {\rm Mpc}^{-1}\Omega_m h^2 = 0.010\ {\rm Mpc}^{-1} \qquad \text{(fiducial cosmology)}. \tag{8.39}$$

In the limit in which we are working, where baryons and anisotropic stresses are neglected, the transfer function depends only on $k/k_{\rm eq}$.

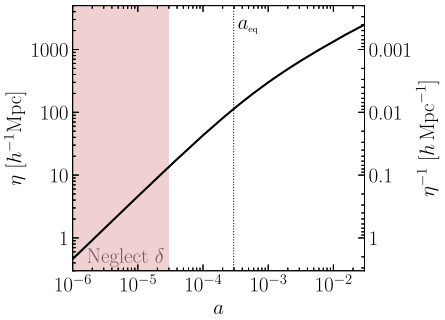



**FIGURE 8.8** The regime studied in Sect. 8.3.1: evolution of modes deep in radiation domination when CDM density perturbations can be neglected.

## 8.3 Small scales

We were able to solve for the evolution of large-scale perturbations in the previous section because the modes crossed the horizon well *after* the epoch of equality. Therefore, the problem is neatly divided into $(i)$ super-horizon modes passing through the epoch of equality and then $(ii)$ modes in the matter-dominated era that cross the horizon. The converse is true for the small-scale modes considered in this section. They cross the horizon when the universe is deep in the radiation era. So the problem divides neatly into $(i)$ modes in the radiation era crossing the horizon (Fig. 8.8) and then $(ii)$ sub-horizon modes passing through the epoch of equality (Fig. 8.12). Step $(i)$ we treat in Sect. 8.3.1, step $(ii)$ in Sect. 8.3.2. Notice that we are unable to treat analytically modes that enter the horizon around the epoch of equality. Nevertheless, nothing is physically different about those modes, they are simply not amenable to the mathematical limits we are taking.

### 8.3.1 Horizon crossing

When the universe is radiation dominated, the potential is determined by perturbations to the radiation. The dark matter perturbations—the ones we are focused on in this chapter—are determined by the gravitational potential, but do not themselves influence the potential. So the situation is as depicted in Fig. 8.9. Solving for matter perturbations in this epoch, therefore, is a two-step problem. First, we must solve the coupled equations for $\Theta_{\rm r,0}$, $\Theta_{\rm r,1}$, and $\Phi$. Then we solve the equation for matter evolution using the potential as an external driving force.

To solve for the potential in the radiation-dominated era, we choose Eq. (8.15). Dropping the matter source terms, we have

$$\Phi = \frac{6a^2H^2}{k^2}\left[\Theta_{\rm r,0} + \frac{3aH}{k}\Theta_{\rm r,1}\right] \tag{8.40}$$

since $H^2 = 8\pi G\rho_{\rm r}/3$ in the radiation era. Also in the radiation era, $aH = 1/\eta$. Armed with this fact, we can use Einstein's equation (8.40) to eliminate $\Theta_{\rm r,0}$ from the two radiation

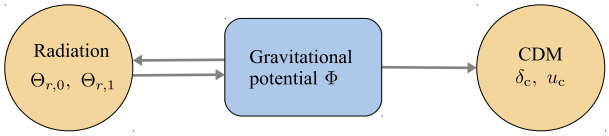



**FIGURE 8.9** Coupling of perturbations during the radiation era. Radiation perturbations and the gravitational potential affect each other. Matter perturbations do not affect the potential but are driven by it.

equations, (8.10) and (8.11). These become

$$-\frac{3}{k\eta}\Theta'_{\mathrm{r},1} + k\Theta_{\mathrm{r},1}\left[1+\frac{3}{k^2\eta^2}\right] = -\Phi'\left[1+\frac{k^2\eta^2}{6}\right] - \Phi\frac{k^2\eta}{3}, \tag{8.41}$$

$$\Theta'_{\mathrm{r},1} + \frac{1}{\eta}\Theta_{\mathrm{r},1} = -\frac{k}{3}\Phi\left[1-\frac{k^2\eta^2}{6}\right]. \tag{8.42}$$

We can turn these two first-order equations for $\Phi$ and $\Theta_{\mathrm{r},1}$ into one second-order equation for the potential. Use Eq. (8.42) to eliminate $\Theta'_{\mathrm{r},1}$ from the first equation, which then becomes

$$\Phi' + \frac{1}{\eta}\Phi = -\frac{6}{k\eta^2}\Theta_{\mathrm{r},1}. \tag{8.43}$$

We now have an expression for $\Theta_{\mathrm{r},1}$ solely in terms of the potential and its first derivative. To arrive at a second-order equation for $\Phi$, we differentiate. When we do so, we will encounter terms proportional to $\Theta_{\mathrm{r},1}$ and its derivative. Each of these can be eliminated with Eq. (8.42) and Eq. (8.43). The resulting second-order equation is

$$\Phi'' + \frac{4}{\eta}\Phi' + \frac{k^2}{3}\Phi = 0. \tag{8.44}$$

This is the wave equation written in Fourier space, with a damping term due to the expansion. Thus, we anticipate oscillating solutions.

To determine the behavior of the potential in the radiation-dominated era, we must solve Eq. (8.44) subject to the initial conditions that $\Phi$ is constant. It can be solved analytically by defining $u \equiv \Phi\eta$. Then Eq. (8.44) becomes

$$u'' + \frac{2}{\eta}u' + \left(\frac{k^2}{3} - \frac{2}{\eta^2}\right)u = 0. \tag{8.45}$$

This is the spherical Bessel equation of order 1 (see Eq. (C.13)) with solutions $j_1(k\eta/\sqrt{3})$ (the spherical Bessel function) and $n_1(k\eta/\sqrt{3})$ (the spherical Neumann function). The latter blows up as $\eta$ approaches zero, so we discard it on the basis of the initial conditions. The spherical Bessel function of order 1 can be expressed in terms of trigonometric functions (Eq. (C.14)), so

$$\Phi(\boldsymbol{k},\eta) = 2\left(\frac{\sin x - x\cos x}{x^3}\right)_{x=k\eta/\sqrt{3}}\mathcal{R}(\boldsymbol{k}). \tag{8.46}$$

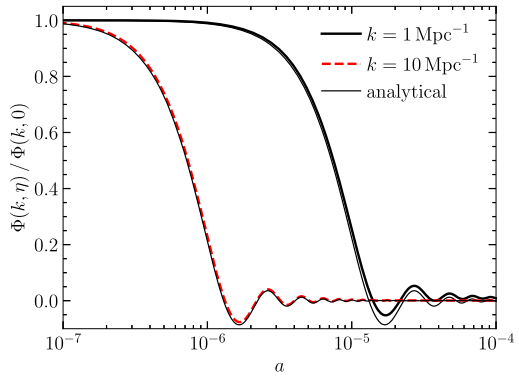



**FIGURE 8.10** Evolution of the potential in the radiation-dominated era. For two small-scale modes which enter the horizon well before equality, the numerical solution is shown along with the approximate analytic solution of Eq. (8.46) (thin lines). We have set the number of neutrinos to zero in the numerical calculation; neutrinos lead to additional damping of the oscillations.

The factor of 2 in front here arises because the $\eta \to 0$ limit of the expression in parentheses is 1/3, and the relation between super-horizon $\Phi$ and $\mathcal{R}$ during radiation domination is $\Phi(\eta \to 0) = 2\mathcal{R}/3$ as we derived in Sect. 7.5.

Eq. (8.46) tells us that, as soon as a mode enters the horizon during the radiation-dominated era, its potential starts to oscillate and decay (Fig. 8.10), as do the density perturbations in the photon-baryon fluid. This is in accord with the qualitative argument at the beginning of this chapter that alluded to the ability of pressure to counteract gravity. Physically, these oscillations are sound waves that are driven by the gravitational potential perturbations as the latter enter the horizon. Recall that we are considering a single Fourier mode here, whose dependence on $\boldsymbol{x}$ is $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, so that the potential induced by a single plane-wave perturbation in Eq. (8.46) is, for $k\eta \gg 1$, roughly described by

$$\Phi(\boldsymbol{x}, \eta) \simeq 6\frac{\mathcal{R}(\boldsymbol{k})}{k^2\eta^2}\cos\left(k\eta/\sqrt{3}\right)\cos(\boldsymbol{k}\cdot\boldsymbol{x}). \tag{8.47}$$

This is the solution for a damped standing wave.

The pressure thus prevents overdensities from growing. If perturbations to the dominant component (here radiation) do not grow, then the potential in an expanding universe will begin to decay. This is evident in Eq. (8.40) which (neglecting the dipole well within the horizon) says that $\Phi \sim \Theta_0/\eta^2$. Since $\Theta_0$ oscillates with fixed amplitude, the potential also oscillates, but with an amplitude decreasing as $\eta^{-2}$. Indeed, this is precisely the large $k\eta$ limit given in Eq. (8.47). The decay and oscillation of the potential is shown in Fig. 8.10 with both the analytic expression of Eq. (8.46) and the numerical solution including matter perturbations. Our approximation, which neglects the effect of dark matter on the potential, leads to deviations that are clearly visible for the large-scale mode: the numerical solution is slightly offset to positive values compared to the analytic result. We will see in Sect. 9.1 that this shift is due to the gravitational effect of dark matter. For the numerical solution

in Fig. 8.10 we have set the number of neutrinos to zero. In the real universe, the presence of free-streaming neutrinos leads to additional damping of the oscillations after horizon entry.

Armed with knowledge of the potential in the radiation-dominated era, we can now determine the evolution of the matter perturbations, the right part of Fig. 8.9. To do this, we turn the two matter evolution equations—Eq. (8.12) and Eq. (8.13)—into one second-order equation with the potentials serving as an external source. Differentiate Eq. (8.12) and use Eq. (8.13) to eliminate $u_c{}'$:

$$\delta_c{}'' + ik\left(-\frac{a'}{a}u_c + ik\Phi\right) = -3\Phi''. \tag{8.48}$$

Now we can use Eq. (8.12) to eliminate $u_c$, leading to

$$\delta_c{}'' + \frac{1}{\eta}\delta_c{}' = S(k,\eta) \tag{8.49}$$

where the source term is

$$S(k,\eta) = -3\Phi'' + k^2\Phi - \frac{3}{\eta}\Phi'. \tag{8.50}$$

The two solutions to the homogeneous equation ($S = 0$) associated with Eq. (8.49) are $\delta_c =$ constant and $\delta_c = \ln(a)$ (or, equivalently in the radiation-dominated era, $\ln(\eta)$). Thus, we anticipate logarithmic growth of $\delta_c$ during the radiation era.

In general, the solution to a second-order equation is a linear combination of the two homogeneous solutions and a particular solution. In the absence of a revelation about the particular solution, one can construct it from the two homogeneous solutions (call them $s_1$ and $s_2$) and the source terms. It is the integral of the source term weighted by the Green function $[s_1(\eta)s_2(\tilde{\eta}) - s_1(\tilde{\eta})s_2(\eta)]/[s_1'(\tilde{\eta})s_2(\tilde{\eta}) - s_1(\tilde{\eta})s_2'(\tilde{\eta})]$. So here we have

$$\delta_c(k,\eta) = C_1 + C_2\ln(k\eta) - \int_0^\eta d\tilde{\eta}S(k,\tilde{\eta})\tilde{\eta}\left(\ln[k\tilde{\eta}] - \ln[k\eta]\right), \tag{8.51}$$

where we have added factors of $k$ in the arguments of the logarithms, which will be convenient later. At very early times the integral is small, so our initial conditions ($\delta_c$ constant) dictate that the coefficient of $\ln(k\eta)$, $C_2$, vanishes and $C_1 = \delta_c(k,\eta = 0) = \mathcal{R}$. Now let us consider the integral in Eq. (8.51). The source function decays to zero along with the potential as the mode enters the horizon. Thus, the dominant contribution to the integral comes from the epochs during which $k\eta$ is of order 1. The integral over $S(\tilde{\eta})\ln(k\tilde{\eta})$ therefore will just asymptote to some constant, while the integral over $S(\tilde{\eta})\ln(k\eta)$ will lead to a term proportional to $\ln(k\eta)$ with the constant of proportionality being just that, a constant. Thus, we expect that after the mode has entered into the horizon,

$$\delta_c(k,\eta) = A\mathcal{R}\ln(Bk\eta), \tag{8.52}$$

i.e., a constant ($A\mathcal{R}\ln[B]$) plus a logarithmic growing mode ($A\mathcal{R}\ln[k\eta]$).

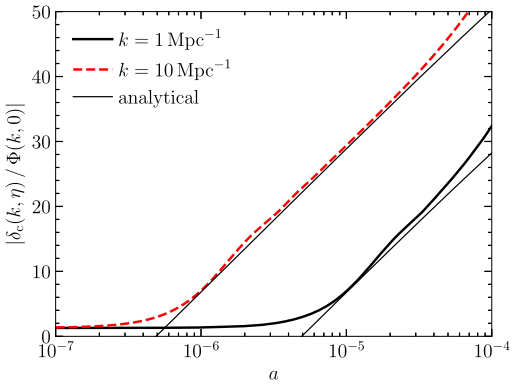



**FIGURE 8.11** Growth of CDM perturbations in the radiation-dominated era. The two modes shown here both enter the horizon in the radiation era and lock onto the logarithmically growing mode. Heavy curves are the numerical solutions, while light solid curves show the logarithmic mode of Eq. (8.52). The perturbations have been normalized by the value of $\Phi$ at early times; the absolute value of the larger-scale mode actually has a larger initial amplitude by a factor of $10^{3/2}$ for a scale-invariant initial spectrum ($n_s = 1$).

We can determine the constants $A$ and $B$ in Eq. (8.52) by referring to the relevant parts of Eq. (8.51). The constant term, $A\mathcal{R}\ln(B)$, is equal to $C_1$ plus the integral over $\ln(\tilde{\eta})$, or

$$A\mathcal{R}\ln(B) = \mathcal{R} - \int_0^\infty d\tilde{\eta} S(k,\tilde{\eta})\tilde{\eta}\ln(k\tilde{\eta}), \tag{8.53}$$

while the coefficient of the $\ln(k\eta)$ term is set by the remaining integral

$$A\mathcal{R} = \int_0^\infty d\tilde{\eta} S(k,\tilde{\eta})\tilde{\eta}. \tag{8.54}$$

Note that in both integrals here, we have set the upper limit to infinity in accord with our expectation that the integrals asymptote to some constant value at large $\eta$. Using the expression for the source term, Eq. (8.50), and our analytic approximations to the potential, Eq. (8.46), we can evaluate the integrals here and determine $A$ and $B$. We find $A = 6.0$ and $B = 0.62$. Hu and Sugiyama (1996), who introduced this method for following the dark matter evolution at early times, found that integrating more precise expressions for the potentials (instead of the approximate ones of Eq. (8.46)) leads to slightly different values, $A = 6.4$ and $B = 0.44$.

Fig. 8.11 shows the numerical solution for $\delta_c$ in the radiation era along with the approximation of Eq. (8.52). Setting aside the details for a moment, we see that matter perturbations do indeed grow even during the radiation era, in contrast to those in the radiation (and baryon) components which decay and oscillate, as we have seen. The reason is that CDM does not have any pressure to counteract the effect of gravity. The growth is not as prominent as during the matter era (when the constant potentials derived in Sect. 8.2 imply $\delta_c \propto a$) due to the more rapid expansion of the universe when radiation dominates, but

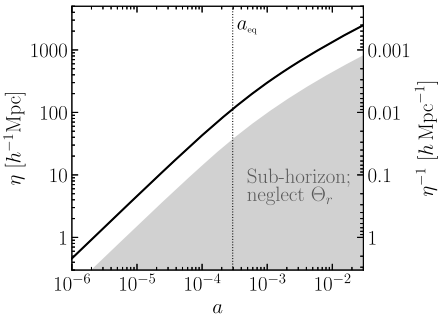



**FIGURE 8.12** The regime studied in Sect. 8.3.2: evolution of modes far inside the horizon, where the effect of radiation perturbations can be neglected.

it still exists. For both modes shown in Fig. 8.11, the perturbations do indeed settle into the logarithmic growing mode once they enter the horizon. As the universe gets closer to matter domination, though, the expansion of the universe slows down, and the perturbations begin to grow faster. Indeed, you might be worried that our approximation for the $k = 1\,h$ Mpc$^{-1}$ mode is not very useful, since the universe enters matter domination soon after it starts to grow. Fortunately, we will be using these solutions only to set the initial conditions for growth in the sub-horizon epoch (next subsection), so the approximation need be valid only over a limited range in time. As long as we choose the matching epoch appropriately, the logarithmic approximation will be quite good.

## 8.3.2 Sub-horizon evolution

We saw in the last subsection that radiation pressure causes the gravitational potentials to decay as modes enter the horizon during the radiation era. Although we did not focus on the radiation perturbations themselves (we will do this in the next chapter), you might expect that the pressure suppresses any growth in $\Theta_{\rm r,0}$. This is correct, and it is in sharp contrast to the matter perturbations which, as we just saw, grow logarithmically. Although initially the potential is determined by the radiation (since the universe is radiation dominated), eventually the growth in the matter perturbations more than offsets the higher mean radiation density. That is, eventually $\rho_{\rm c}\delta_{\rm c}$ becomes larger than $\rho_{\rm r}\Theta_{\rm r,0}$ even if $\rho_{\rm c}$ is smaller than $\rho_{\rm r}$. Once this happens, the gravitational potential and the dark matter perturbations evolve together and do not care what happens to the radiation (Fig. 8.12). In this subsection, we want to solve the set of equations governing the matter perturbations and the potential and then match on to the logarithmic solution (8.52) set up during the epoch in which the potential decays.

Once again our starting point is the set of equations governing dark matter evolution, Eq. (8.12) and Eq. (8.13), and the algebraic equation for the gravitational potential (8.15). And, once again, we want to reduce this set of three equations (two of which are first-order differential equations) to one second-order equation. We will want to follow the sub-horizon dark matter perturbations through the epoch of equality, so it proves con-

venient again to use $y$ (Eq. (8.20))—the ratio of the scale factor to its value at equality—as the evolution variable. In terms of $y$, the three equations become

$$\frac{d\delta_c}{dy} + \frac{iku_c}{aHy} = -3\frac{d\Phi}{dy}, \tag{8.55}$$

$$\frac{du_c}{dy} + \frac{u_c}{y} = \frac{ik\Phi}{aHy}, \tag{8.56}$$

$$k^2\Phi = \frac{3y}{2(y+1)}a^2H^2\delta_c. \tag{8.57}$$

Several comments are in order about this version of our fundamental equations. First, notice that the time derivatives in the first two equations have been replaced by derivatives with respect to $y$, and this transformation leads to the factors of $y' = aHy$ in the denominators of the other terms. Second, the gravitational potential is now expressed solely in terms of $\delta_c$: there is no dependence on radiation perturbations because of our arguments above that these are subdominant, and there is no $aHu_c/k$ dependence because the perturbations are well within the horizon and $aH/k \ll 1$. Finally, the coefficient of the $\delta_c$ source term is $4\pi G\rho_c a^2 \to (3/2)a^2H^2y/(y+1)$. Here we neglect both baryons and dark energy. The latter is a good assumption since we are interested in times early enough that the effect of dark energy is negligible; the former will lead to some differences from the correct numerical result.

We now go through the familiar routine of turning Eqs. (8.55) and (8.56) into a second-order equation for $\delta_c$: differentiate the first of these with respect to $y$ to get

$$\frac{d^2\delta_c}{dy^2} - \frac{ik(2+3y)u_c}{2aHy^2(1+y)} = -3\frac{d^2\Phi}{dy^2} + \frac{k^2\Phi}{a^2H^2y^2} \tag{8.58}$$

where $du_c/dy$ has been eliminated using the velocity equation. Also we have used the fact that $d(1/aHy)/dy = -(1+y)^{-1}(2aHy)^{-1}$. The first term on the right is much smaller than the second, which is multiplied by $(k/aH)^2$, and can thus be dropped. Using Eq. (8.57), we recognize this second term as $3\delta_c/[2y(y+1)]$. We can rewrite the velocity on the left using Eq. (8.55) but neglecting the potential which on sub-horizon scales is much smaller than $\delta_c$. Thus, the combination $iku_c/(aHy)$ can be simply replaced by $-d\delta_c/dy$ leaving

$$\frac{d^2\delta_c}{dy^2} + \frac{2+3y}{2y(y+1)}\frac{d\delta_c}{dy} - \frac{3}{2y(y+1)}\delta_c = 0. \tag{8.59}$$

This is the *Meszaros equation* (Meszaros, 1974), governing the evolution of sub-horizon cold dark matter perturbations once radiation perturbations have become negligible.

To understand the growth of dark matter perturbations, we need to obtain the two independent solutions to the Meszaros equations and then match on to the logarithmic mode established in the previous subsection. To solve this differential equation, we can use our knowledge of the solution deep in the matter era. We have seen that sub-horizon perturbations in the matter era grow with the scale factor (and will prove this in Sect. 8.5), so

one of the solutions to Eq. (8.59) is a polynomial in $y$ of order 1. Therefore, for one mode at least, $d^2\delta_c/dy^2$ vanishes. The equation governing this first mode, the growing mode, is $\delta'_{c,+}/\delta_{c,+} = 3/(2+3y)$, the solution to which is $\delta_{c,+} \propto y + 2/3$, or

$$D_+(a) = a + \frac{2a_{\rm eq}}{3}. \tag{8.60}$$

Normalized in this way, the solution describes scale-independent growth, and approaches $D_+ = a$ when $a \gg a_{\rm eq}$. Hence, we have identified it with the *growth factor* $D_+$ introduced in Sect. 8.1. Note though that in this section we are assuming that only matter is relevant, and ignore curvature and dark energy. Therefore, our expression for the growth factor will be valid only when $a \lesssim 0.1$. We turn to the generalization to later times in Sect. 8.5.

To find the second solution, notice that the Meszaros equation tells us that $u \equiv \delta_c/(y + 2/3)$ satisfies

$$(1+3y/2)\frac{d^2u}{dy^2} + \frac{(21/4)y^2 + 6y + 1}{y(y+1)}\frac{du}{dy} = 0. \tag{8.61}$$

Since there is no term proportional to $u$, Eq. (8.61) is actually a first-order equation for $du/dy$. We can therefore integrate to obtain a solution for $du/dy$ and then integrate again to get the second Meszaros solution. The first integral gives

$$\frac{du}{dy} \propto (y+2/3)^{-2} y^{-1} (y+1)^{-1/2}. \tag{8.62}$$

Integrating again leads to the second Meszaros solution

$$D_-(y) = (y+2/3)\ln\left[\frac{\sqrt{1+y}+1}{\sqrt{1+y}-1}\right] - 2\sqrt{1+y}. \tag{8.63}$$

At early times ($y \ll 1$), $D_+ = $ const while $D_-$ is proportional to $\ln y$. At late times ($y \gg 1$), the growing solution $D_+$ scales as $y$ while the decaying mode $D_-$ falls off as $y^{-3/2}$.

The general solution to the Meszaros equation is therefore

$$\delta_c(k, y) = C_1 D_+(y) + C_2 D_-(y) \qquad (y \gg y_H) \tag{8.64}$$

where $y_H \equiv a_H/a_{\rm eq}$ is the scale factor when the mode enters the horizon divided by the scale factor at equality (Exercise 8.6). To determine the constants $C_1$ and $C_2$ we can match on to the logarithmic solution of Eq. (8.52). That solution is valid within the horizon but before equality: $y_H \ll y \ll 1$. So we can hope to arrive at a reasonable approximation for the evolution of dark matter perturbations only for those modes that enter the horizon before equality. For those modes, we match the two solutions and their first derivatives,

$$\begin{aligned} A\mathcal{R}\ln(By_m/y_H) &= C_1 D_+(y_m) + C_2 D_-(y_m), \\ \frac{A\mathcal{R}}{y_m} &= C_1 D'_+(y_m) + C_2 D'_-(y_m), \end{aligned} \tag{8.65}$$

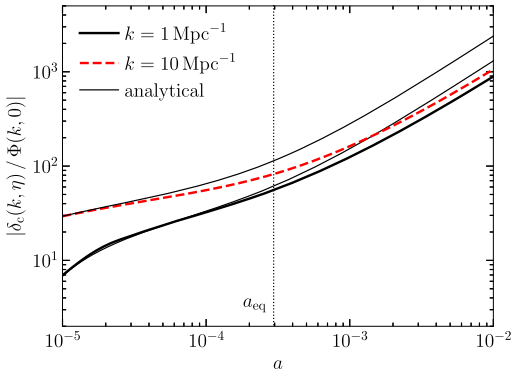



**FIGURE 8.13** Evolution of small-scale, sub-horizon, dark matter perturbations. Thick curves are numerical solutions; thin curves show the Meszaros solution with coefficients given by the matching condition, Eq. (8.65), applied at $y_m = 3y_H$. The departure from the numerical solution at late times is due to the presence of baryons.

where the matching epoch $y_m$ must satisfy $y_H \ll y_m \ll 1$. Note that we have replaced the argument $k\eta$ of the log in Eq. (8.52) with $y/y_H$, which is valid as long as the matching epoch is deep in the radiation era. Fig. 8.13 shows the evolution of two modes along with the analytic solutions to the Meszaros equation with coefficients set by the matching conditions given in Eq. (8.65). The departure at $a > a_{\rm eq}$ is due to the presence of baryons which we have neglected here. After equality, the contribution of baryons to the gravitational potential is not negligible, but they cluster less than the dark matter due to their coupling to photons up until recombination. Since in our calculation we assumed that all matter is in CDM form, our analytic solution overestimates the growth of the CDM component.

## 8.4 The transfer function

In Sect. 8.2 and Sect. 8.3, we derived analytic solutions following the dark matter perturbations deep into the matter era. Here, we assemble these results to obtain an idea of the form of the transfer function.

First, we need to transform our expression (Eq. (8.64) along with Eq. (8.65)) for the small-scale matter density into an expression for the transfer function. The transfer function is determined by the behavior of $\delta_{\rm c}$ well after equality when the decaying mode has long since vanished. We can extract an even simpler form for $\delta_{\rm c}$ in this $a \gg a_{\rm eq}$ limit. The key constant in that case is $C_1$, the coefficient of the growing mode. Multiplying the first matching condition in Eq. (8.65) by $D'_-$ and the second by $D_-$ and then subtracting lead to

$$C_1 = \frac{D'_-(y_m)\ln(By_m/y_H) - D_-(y_m)/y_m}{D_+(y_m)D'_-(y_m) - D'_+(y_m)D_-(y_m)}A\mathcal{R}. \tag{8.66}$$

The denominator $D_+D'_- - D'_+D_- = -(4/9)y_m^{-1}(y_m+1)^{-1/2}$, which is approximately equal to $-4/9y_m$ since $y_m \ll 1$. Similarly for small $y_m$, $D_- \to (2/3)\ln(4/y) - 2$ and $D'_- \to -2/3y$.

Therefore,

$$C_1 \to -\frac{9}{4}A\mathcal{R}\left[\frac{-2}{3}\ln(By_m/y_H) - (2/3)\ln(4/y_m) + 2\right], \tag{8.67}$$

which fortuitously does not depend on $y_m$. Therefore, at late times we have an approximate solution for the small-scale dark matter perturbations

$$\delta_{\rm c}(\boldsymbol{k}, a) = \frac{3}{2}A\mathcal{R}(\boldsymbol{k})\ln\left[\frac{4Be^{-3}a_{\rm eq}}{a_H}\right]D_+(a) \qquad (a \gg a_{\rm eq}). \tag{8.68}$$

On very small scales, the argument of the log simplifies because $a_{\rm eq}/a_H = \sqrt{2}k/k_{\rm eq}$ (Exercise 8.6). We also need to remember that we have been ignoring baryons throughout, and so within this approximation we set the total matter density perturbation to that of CDM: $\delta_{\rm m} = \delta_{\rm c}$ (we return to this in Sect. 8.6.1). Comparing Eq. (8.7) with Eq. (8.68) leads to an analytic expression for the transfer function on small scales:

$$T(k) = \frac{15}{4}\frac{\Omega_{\rm m}H_0^2}{k^2 a_{\rm eq}}A\ln\left[\frac{4Be^{-3}\sqrt{2}k}{k_{\rm eq}}\right] \qquad (k \gg k_{\rm eq}). \tag{8.69}$$

Recall that the wavenumber entering the horizon at equality is defined as

$$k_{\rm eq} \equiv a_{\rm eq}H(a_{\rm eq}) = \sqrt{2\Omega_{\rm m}}H_0 a_{\rm eq}^{-1/2}, \tag{8.70}$$

so the prefactor is also a function of $k/k_{\rm eq}$ only. Then, plugging in numbers ($A = 6.4$, $B = 0.44$) leads to

$$T(k) = 12.0\frac{k_{\rm eq}^2}{k^2}\ln\left[0.12\frac{k}{k_{\rm eq}}\right] \qquad (k \gg k_{\rm eq}). \tag{8.71}$$

This analytic approximation for the transfer function is only accurate on very small scales, $k \gtrsim 1\,h\,{\rm Mpc}^{-1}$. More sophisticated analytic solutions or fitting formulae have lost most of their practical utility since the advent of fast and accurate codes to compute the transfer function (these codes now agree to within 0.1%). Importantly though, our analytic work has enabled us to understand the origin of the asymptotic, small-scale behavior of the transfer function. Had there been no logarithmic growth in the radiation era, the modes that entered very early on would have experienced no growth from horizon entry until the epoch of equality. Their amplitude relative to large-scale modes would then have been suppressed by a factor of order $(k_{\rm eq}/k)^2$. The logarithmic growing mode in the radiation era ameliorates this suppression (the effect is seen in the larger amplitude of the higher-$k$ mode in Fig. 8.13).

Finally, we can look at the shape of the matter power spectrum via Eq. (8.8). Fig. 8.14 shows the power spectrum for our fiducial Euclidean $\Lambda$CDM cosmology, as well as the effect of changing $\Omega_{\rm m}$ while retaining a Euclidean cosmology by requiring $\Omega_{\rm m} + \Omega_\Lambda = 1$ and keeping $h$ fixed. Clearly, the shape of the power spectrum, and in particular the turnover

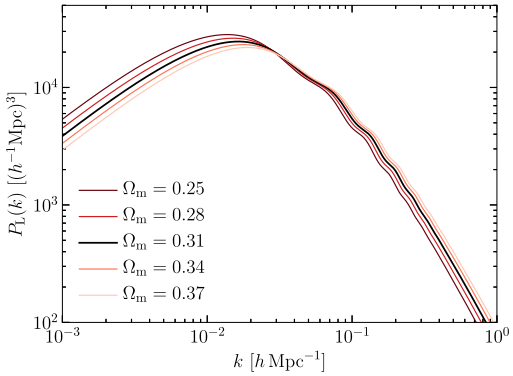



**FIGURE 8.14** The matter power spectrum at redshift $z = 0$ in the fiducial $\Lambda$CDM cosmology (thick black line). The other lines show the result when varying $\Omega_\mathrm{m}$ around the fiducial value, keeping $h$ fixed and $\Omega_\mathrm{m} + \Omega_\Lambda = 1$. Changing $\Omega_\mathrm{m}$ changes the epoch of equality and hence the shape of the matter power spectrum.

scale $k_\mathrm{eq}$ depend on $\Omega_\mathrm{m}$: as we lower $\Omega_\mathrm{m}$, the matter-radiation equality shifts to later times, so that $k_\mathrm{eq}$ is pushed to lower values; the opposite happens when increasing $\Omega_\mathrm{m}$. Recall that the physical matter density is controlled by the parameter combination $\Omega_\mathrm{m}h^2$. Hence, $k_\mathrm{eq}$ and the shape of the matter power spectrum mainly depend on this combination, as you will verify in Exercise 8.7. There is another subtlety, however, which is purely conventional: in order to make distances independent of the Hubble constant, length scales are conventionally multiplied by $h$, and wavenumbers correspondingly divided by $h$ (see the axis labels in Fig. 8.14 which indicate this). Since $k_\mathrm{eq}$ is proportional to $\Omega_\mathrm{m}h^2$ in physical units (i.e., $\mathrm{Mpc}^{-1}$, cf. Eq. (8.39)), the parameter combination controlling $k_\mathrm{eq}$ in units of $h\,\mathrm{Mpc}^{-1}$ is $\Omega_\mathrm{m}h$. The combination $\Omega_\mathrm{m}h$ is thus sometimes referred to as "shape parameter." It is important to keep in mind, however, that in terms of physical scales, the relevant parameter is $\Omega_\mathrm{m}h^2$.

There are several physical effects in the real universe that we have neglected in our analytic treatment. We have assumed no anisotropic stress ($\Phi = -\Psi$). Dropping this assumption changes the factor of 9/10 by which the potential drops for large-scale modes to approximately 0.86, resulting in a corresponding rise in the small-scale transfer function. Including the effect of baryons leads to even more significant small-scale changes. We will address these in Sect. 8.6. Third, all of our work in this section has been on the transfer function, i.e., on the evolution of perturbations early on when the only relevant constituents of the universe were matter and radiation. At late times, the growth factor depends on other constituents, the most important of which is dark energy.

## 8.5 The growth factor

Armed with knowledge of the transfer function, let us now turn to the second part of structure formation, the scale-independent growth factor. At late times, the horizon is much larger than the modes that interest us. Were it not for dark energy (and neutrino masses), we could simply continue to use the Meszaros equation of Sect. 8.3.2.

In order to begin our calculation, let us then recap what regime we will focus on. Apart from the sub-horizon limit, we can also neglect the pressure in the baryons at late times, i.e. after decoupling. This means that the baryons follow equations that look just like those for the dark matter, Eqs. (8.12)–(8.13). Moreover, while they start with different initial conditions, the baryons closely follow the dark matter at late times, so we can describe the matter sector with the total matter perturbation, defined through (Eq. (6.79)) $\rho_{\rm m}\delta_{\rm m} = \rho_{\rm c}\delta_{\rm c} + \rho_{\rm b}\delta_{\rm b}$, and similarly for the velocity $u_{\rm m} = (\rho_{\rm c}u_{\rm c} + \rho_{\rm b}u_{\rm b})/\rho_{\rm m}$. In this section, we will also neglect the mass of neutrinos, whose effect does complicate the late-time evolution of structure.

First, multiply Eq. (8.12) by $a$ and take the derivative with respect to $\eta$. Neglecting the right-hand side, which is negligible on sub-horizon scales, and combining with Eq. (8.13) yields

$$\left[a\delta_{\rm m}'(\boldsymbol{k},\eta)\right]' = ak^2\Phi(\boldsymbol{k},\eta). \tag{8.72}$$

We now need to complement this with one of the Einstein equations for $\Phi$. Let us take Eq. (8.14), neglect radiation perturbations as well as the terms on the left-hand side that are small when $k \gg aH$. We obtain

$$k^2\Phi(\boldsymbol{k},\eta) = 4\pi G a^2\rho_{\rm m}(\eta)\delta_{\rm m}(\boldsymbol{k},\eta), \tag{8.73}$$

which is nothing but Eq. (8.5). Using the fact that $\rho_{\rm m} \propto a^{-3}$ and the definition of $\Omega_{\rm m}$, we finally obtain the first version of our growth equation for $\delta_{\rm m}$:

$$\left[a\delta_{\rm m}'\right]' = \frac{3}{2}\Omega_{\rm m}H_0^2\delta_{\rm m}. \tag{8.74}$$

For solving this equation, it is more convenient to exchange the time variable from $\eta$ to $a$, which yields

$$\frac{d^2\delta_{\rm m}}{da^2} + \frac{d\ln(a^3H)}{da}\frac{d\delta_{\rm m}}{da} - \frac{3\Omega_{\rm m}H_0^2}{2a^5H^2}\delta_{\rm m} = 0. \tag{8.75}$$

In general, Eq. (8.75) needs to be solved numerically. There are a few (important) special cases where we can make a bit more progress. You will show in Exercise 8.8 that, if the only relevant components apart from matter are a cosmological constant and curvature, we can obtain the following integral solution:

$$D_+(a) \propto H(a)\int^a \frac{da'}{(a'H(a'))^3} \qquad (\Lambda,\ \text{curvature})\,. \tag{8.76}$$

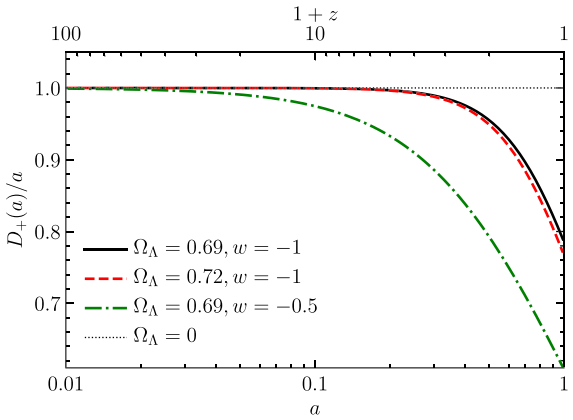



**FIGURE 8.15** The growth factor divided by the scale factor in three different Euclidean cosmologies. The solid line shows the fiducial cosmology. Increasing the amount of dark energy, or increasing its equation of state above $-1$, leads to a greater suppression of the growth at late times.

The proportionality constant is fixed by the definition of Eq. (8.3), which says that, early on when matter still dominates (say at $z \simeq 10$), $D_+$ should be equal to $a$. At those times, $H = H_0 \Omega_{\rm m}^{1/2} a^{-3/2}$, so the growth factor is

$$D_+(a) = \frac{5\Omega_{\rm m}}{2} \frac{H(a)}{H_0} \int_0^a \frac{da'}{(a' H(a')/H_0)^3} \qquad (\Lambda,\ \text{curvature})\,. \tag{8.77}$$

If dark energy is not a cosmological constant, then Eq. (8.77) is *not* a solution to the second-order growth equation, which needs to be solved directly. However, for the *growth rate* $f$, the logarithmic derivative of the growth factor, there exists an empirical fitting formula that is remarkably precise even in the presence of dynamical dark energy:

$$f(a) \equiv \frac{d\ln D_+(a)}{d\ln a} \simeq [\Omega_{\rm m}(a)]^{0.55}\,, \tag{8.78}$$

where $\Omega_{\rm m}(a) \equiv 8\pi G\rho_{\rm m}(a)/3H^2(a)$ is the time-dependent matter density parameter (which reduces to our constant $\Omega_{\rm m}$ if $a = 1$). We will use this time-dependent $\Omega_{\rm m}(a)$ only here and in Ch. 12.

Fig. 8.15 shows the growth factor for three different Euclidean cosmologies, divided by $a$ in order to better show the trends at late times. As mentioned above, if the universe is Euclidean and matter dominated, the growth factor is simply equal to the scale factor. In the presence of dark energy, growth is suppressed by varying amounts depending on the amount and equation of state of dark energy. We will see some observable implications of this in Ch. 11.

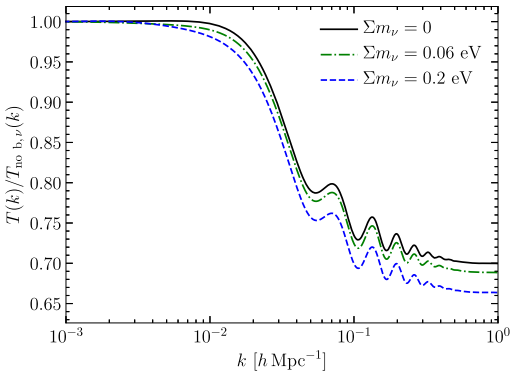



**FIGURE 8.16** Ratio of the transfer function at redshift zero to the case of no baryons (i.e. $\Omega_c = \Omega_m$) and massless neutrinos. The black solid line shows the effect of baryons, which is very conspicuous. The other lines show the additional effect of finite neutrino masses. Here, we have assumed equal masses for all three neutrino species. While the effect of baryons is essentially independent of redshift, massive neutrinos lead to a redshift dependence of the transfer function.

## 8.6 Beyond cold dark matter and radiation

Although CDM is the main matter component, so that the transfer function we derived earlier is a reasonable approximation to reality, there is more to the real universe than just cold dark matter. Here we focus on three additional components. First, we consider the effect of the baryons, which constitute roughly 16% of the total matter, on the transfer function. Then, we study the consequences of the fact that neutrinos have mass. Finally, dark energy—one model for which is the cosmological constant—is considered.

### 8.6.1 Baryons

A careful examination of the black solid line in Fig. 8.16 reveals two signatures of the presence of baryons. The first is that the transfer function is suppressed relative to the no-baryon case on small scales. This is not surprising: at early times, before decoupling, baryons are tightly coupled to photons. Since radiation perturbations do not grow inside the horizon, the baryon overdensities are likewise suppressed compared to the dark matter. After decoupling, baryons are released from the relatively smooth radiation field and fall into the gravitational potentials dominated by the dark matter. The depth of these wells is smaller than we estimated in Sect. 8.3, though, because only a fraction $\Omega_c/\Omega_m$ of the total matter was involved in the collapse.

The second effect of baryons visible in Fig. 8.16 is equally important. Baryons lead to small oscillations in the transfer function around $k \simeq 0.1\,h\,\mathrm{Mpc}^{-1}$. These are manifestations of the oscillations (sound waves) that the combined baryon-photon fluid experiences before decoupling. We encountered these in Sect. 8.3.2 (Fig. 8.10) when we derived the potential in the radiation-dominated era. The oscillations in the potential reflect those in the density of the baryon-photon fluid, which are acoustic plasma waves. For this reason, the

oscillations in the transfer function are known as *baryon acoustic oscillations (BAOs)*. They have been detected with impressive significance in the clustering of galaxies (Fig. 1.9). Since this feature corresponds to a known scale, the sound horizon at decoupling (roughly, $\eta_*/\sqrt{3}$, cf. Eq. (8.47)), it can be used as a standard ruler if measured in the large-scale structure of the late universe as shown in Fig. 1.9.

The main difficulty in measuring this feature is its small amplitude, which is simply due to the fact that baryons are such a small fraction of the total matter. The oscillations are much more prominent in the radiation, as we already saw in Fig. 1.10. How to describe those accurately is the topic of the next chapter.

Now, above in Sect. 8.5 we argued that baryons eventually follow the dark matter after decoupling. It is worth investigating in a little more detail how this happens. Let us go back to Eqs. (8.12)–(8.13). After decoupling, the baryons are free of their coupling to photons and have negligible pressure since their temperature is low. Hence, they obey the same equations as the CDM component, and we arrive at the following set of equations:

$$\begin{aligned} {\delta_s}' + iku_s &= -3\Phi', \\ {u_s}' + \frac{a'}{a}u_s &= ik\Phi \qquad (s = \{\mathrm{b},\, \mathrm{c}\}). \end{aligned} \tag{8.79}$$

What we did in Sect. 8.5 is to construct the weighted sums

$$\begin{aligned} \delta_\mathrm{m} &= \frac{\rho_\mathrm{c}\delta_\mathrm{c} + \rho_\mathrm{b}\delta_\mathrm{b}}{\rho_\mathrm{m}}, \\ u_\mathrm{m} &= \frac{\rho_\mathrm{c}u_\mathrm{c} + \rho_\mathrm{b}u_\mathrm{b}}{\rho_\mathrm{m}}, \end{aligned} \tag{8.80}$$

i.e. the total matter density perturbation and matter velocity. Combining this with the Poisson equation, which is sourced by $\delta_\mathrm{m}$, then yields Eq. (8.75).

Now let us construct a different set of variables, the relative density perturbation and relative velocity between baryons and CDM:

$$\begin{aligned} \delta_\mathrm{bc} &= \delta_\mathrm{b} - \delta_\mathrm{c}, \\ u_\mathrm{bc} &= u_\mathrm{b} - u_\mathrm{c}. \end{aligned} \tag{8.81}$$

We obtain equations for these by subtracting the continuity and velocity equations for baryons and CDM:

$$\begin{aligned} {\delta_\mathrm{bc}}' + iku_\mathrm{bc} &= 0, \\ {u_\mathrm{bc}}' + \frac{a'}{a}u_\mathrm{bc} &= 0. \end{aligned} \tag{8.82}$$

Notice that the gravitational potential drops out of both equations. This is due to the fact that gravity cares only about the total amount of matter, not how much of it is in baryons or dark matter. The two solutions to Eq. (8.82) are easily identified: first, we have $\delta_\mathrm{bc} = C_\delta$,

$u_{\rm bc} = 0$. This corresponds to a constant relative density perturbation between baryons and CDM, keeping the total amount of matter fixed, and with both traveling with the same velocity. Second, the equation for the relative velocity admits a solution $u_{\rm bc} = C_u/a$, with $\delta_{\rm bc} \propto C_u \int d\eta/a$. This solution corresponds to giving the baryons an initial push, so that they acquire a relative velocity with respect to the dark matter. Both of these modes are generated by the different evolution of baryons and dark matter leading up to decoupling. The key feature though is that these modes are constant or decaying, and thus become small at late times compared to the growing mode studied in Sect. 8.5. That is, even if at decoupling $\delta_{\rm bc}$ is comparable to $\delta_{\rm m}$, it is suppressed by a factor $D(a_*)/D(a) \lesssim 0.01$ in the late universe; the suppression of $u_{\rm bc}$ is even stronger.

### 8.6.2 Massive neutrinos

Neutrinos are known to exist, and the concordance model of cosmology makes a definite prediction for how many there are in the universe (Eq. (2.82)); only their mass remains uncertain. An accurate measurement of the matter power spectrum may enable us to infer neutrino masses.

The masses of neutrinos affect structure growth in two ways. First, they affect the evolution of the energy density in neutrinos, which initially decays as $a^{-4}$ but later transitions to $a^{-3}$ (cf. Eq. (2.83)). This modifies the expansion rate through the Friedmann equation, and changes the growth factor since $H(a)$ enters the growth equation (8.75). Taking into account this effect, which you can do in Exercise 8.10, still leaves the growth factor independent of $k$.

The second effect is a bit more subtle. Since neutrinos move fast (they are not *cold* dark matter) and stream out of high-density regions, they damp the growth of small-scale structure. Perturbations on scales smaller than the typical distance that neutrinos travel, the *free-streaming scale*, are therefore suppressed. We can estimate the scale on which perturbations are damped by computing the comoving distance a massive neutrino can travel in one Hubble time. As you can show in Exercise 8.11, the inverse of this, the free-streaming wavenumber, is given by

$$k_{\rm fs}(a) \simeq 0.063\, h\ {\rm Mpc}^{-1}\, \frac{m_\nu}{0.1\ {\rm eV}}\, \frac{a^2 H(a)}{H_0}. \tag{8.83}$$

Fig. 8.16 shows the resulting suppression (note that it does not include the first neutrino-mass effect, the modified expansion history, which is not included in $T(k)$ by definition). More massive neutrinos constitute more of the total density so they suppress small-scale power more than do lighter neutrinos. However, there is a smaller effect, barely noticeable for $k \lesssim 0.004\, h$ Mpc$^{-1}$ in Fig. 8.16, that works in the other direction. Less massive neutrinos travel more rapidly and hence free-stream out of larger regions, as indicated by Eq. (8.83). Therefore the transfer function for $\sum m_\nu = 0.06$ eV is slightly smaller than that for $\sum m_\nu = 0.2$ eV on very large scales.

Finally, notice that the growth factor becomes scale dependent in the presence of massive neutrinos, i.e. the neat decomposition into a time-independent transfer function and scale-independent growth factor is upset.

### 8.6.3 Dark energy

As reviewed in Sect. 2.4.6, we now have overwhelming evidence that dark energy dominates the energy budget of the universe today. How does it affect the matter perturbations?

The direct physical effect of dark energy is the impact on the growth factor we derived in Sect. 8.5. Following Eq. (8.77), the growth of perturbations at late times depends directly on the evolution of the Hubble rate, which in turn depends on the amount and evolution of dark energy. Thus, different models of dark energy predict different growth factors. If we parameterize the dark energy by its equation of state $w$ (Eq. (2.60)), and assume that $w$ is constant, then the Hubble rate in a Euclidean universe evolves as

$$\frac{H(z)}{H_0} = \left[\frac{\Omega_\mathrm{m}}{a^3} + \frac{\Omega_\mathrm{de}}{a^{3(1+w)}}\right]^{1/2} \tag{8.84}$$

at late times. Using this time dependence, it is straightforward to solve Eq. (8.75) numerically (Exercise 8.9; notice that Eq. (8.77) is not valid if $w \neq -1$; see Exercise 8.8). This effect, in addition to the distance-redshift relation discussed in Sect. 2.2, forms the basis for current constraints on the dark energy density and equation of state from large-scale structure.

Two additional, indirect effects of dark energy enter through the fact that, in a Euclidean universe, $\Omega_\mathrm{m} = 1 - \Omega_\Lambda$. First, the shape of the matter power spectrum indirectly depends on the amount of dark energy since the equality scale depends on $\Omega_\mathrm{m}$. Second, for a fixed amplitude of potential perturbations, which is what the large-angle CMB anisotropies constrain, the fractional density perturbations are proportional to $\Omega_\mathrm{m}^{-1}$, which follows from the Poisson equation (8.6).

## 8.7 Summary

After having set up the relevant linear Einstein–Boltzmann system and the initial conditions from inflation in previous chapters, we were now able to derive solutions to these equations, focusing on the growth of density perturbations $\delta_\mathrm{c}$ in the dark matter. These form the basis for the structure in the late universe, and are thus well worth studying.

The focus on dark matter allowed us to strongly simplify the treatment of baryons and radiation, and we obtained analytic results in important special cases: large-scale modes that entered the horizon during matter domination, and small-scale modes that entered the horizon deep in radiation domination. During radiation domination, perturbations in the radiation oscillate in the form of standing sound waves, while those in dark matter grow logarithmically. Later, during matter domination, dark matter perturbations grow proportionally to the scale factor.

Apart from these limits which can be solved analytically, the growth of structure in general has to be integrated numerically. This can be done efficiently using publicly available codes. The two most widely used codes for computing transfer functions, which also output the matter power spectrum and angular power spectra of CMB anisotropies, are *CAMB* (Lewis et al., 2000) and *CLASS* (Blas et al., 2011). The plots in this chapter were made using CLASS via its python module.

Because structure growth becomes scale-free after decoupling, it is convenient to decompose the growth into a scale-dependent *transfer function* $T(k)$ and a time-dependent *growth factor* $D_+(a)$. This parametrization neatly decouples early-universe physics—contained in $T(k)$—from late-time physics such as dark energy—captured by $D_+(a)$. Using these definitions, the matter density at late times is related to the conserved curvature perturbation generated by inflation via

$$\delta_{\rm m}(\boldsymbol{k}, a) = \frac{2}{5}\frac{k^2}{\Omega_{\rm m} H_0^2}\mathcal{R}(\boldsymbol{k})T(k)D_+(a) \qquad (a > a_{\rm late},\ k \gg aH). \tag{8.85}$$

The main result of this chapter was our prediction for the matter power spectrum, which becomes

$$P_{\rm L}(k, a) = \frac{8\pi^2}{25}\frac{\mathcal{A}_s}{\Omega_{\rm m}^2} D_+^2(a) T^2(k) \frac{k^{n_s}}{H_0^4 k_{\rm p}^{n_s-1}}. \tag{8.86}$$

It is worth noting that massive neutrinos spoil the decomposition into transfer function and growth factor on scales smaller than the free-streaming scale, i.e. for $k \gtrsim k_{\rm fs}$: $D_+$ becomes scale dependent, or equivalently $T(k)$ becomes time dependent.

While computing an accurate transfer function requires substantial effort in terms of equations and physics, the growth factor (at least on scales larger than the free-streaming scale $1/k_{\rm fs}$) can be computed fairly easily as soon as the expansion history is specified, by solving the ordinary differential equation Eq. (8.75) with initial conditions given by $D_+(a) = a$ during matter domination.

With these ingredients, we are all set to compare the matter power spectrum with observations of large-scale structure, which we will do in Ch. 11. Implicitly, we normalized the power spectrum to large-scale CMB anisotropies in Eq. (8.86); we will understand how this works in the next chapter. The fact that the matter power spectrum determined by theory together with early-universe measurements agrees with the completely independent measurements in the late universe is a highly nontrivial consistency test of our concordance cosmological model.

## Exercises

**8.1** Derive Eqs. (8.10) and (8.11).

**(a)** Show that, in the limit of small baryon density, the scattering term in Eq. (5.67), the one proportional to $\tau'$, can be neglected: first, drop $\Pi$, since the quadrupole

and polarization are very small. Then show that the scattering term is proportional to the baryon-to-photon energy ratio $R$ defined in Eq. (5.74). You will want to use Eq. (5.72). Again, this series of approximations is valid only for the purposes of this chapter, wherein we are interested in the matter distribution.

**(b)** Neglecting the scattering term in Eq. (5.67), show that this collisionless equation reduces to the two equations for the monopole and dipole. To get the monopole equation, multiply Eq. (5.73) by $\mathcal{P}_0(\mu) = 1$ and integrate over $d\mu/2$. To get the dipole, multiply by $\mathcal{P}_1(\mu)$ and integrate.

**8.2** Solve the set of five equations (Eq. (8.10)–Eq. (8.13) and Eq. (8.14)) numerically to obtain the transfer function for dark matter. Use the initial conditions derived in Ch. 7. The one numerical problem you may encounter using Eq. (8.14) occurs on small scales when you try to evolve all the way to the present. The photon moments then become difficult to track, and even a good differential equation solver will balk at late times. However, there are several simple solutions to this: $(i)$ by the late times in question, the potential is constant so there is no need to evolve all the way to the present or $(ii)$ stop following the photon moments after a certain time; they do not have any effect on the matter distribution at late times anyway. Plot the transfer function for the fiducial ΛCDM model and compare with the output of CAMB or CLASS.

**8.3** The four subsections in Sect. 8.2 and Sect. 8.3 correspond to four different approximations to the full set of Einstein–Boltzmann equations. In the following table, fill in the regime of validity for each approximation:

| | $a \ll a_{\rm eq}$ | $a \sim a_{\rm eq}$ | $a \gg a_{\rm eq}$ |
|---|---|---|---|
| $k\eta \ll 1$ | | | |
| $k\eta \sim 1$ | | | |
| $k\eta \gg 1$ | | | |

For example, the super-horizon solution of Sect. 8.2.1 is valid along the whole top row, since it sets $k\eta \to 0$. Note that time evolves from upper left to bottom right, so the fact that none of the approximations work in the center square means that only those scales that enter the horizon well before or well after equality will be amenable to analytic techniques.

**8.4** Fill in some of the algebraic detail left out of Sect. 8.2.1.

**(a)** Show that Eq. (8.23) leads to Eq. (8.24) by carrying out the differentiation.

**(b)** Show that Eq. (8.24) is equivalent to Eq. (8.26) when the definition of $u$ from Eq. (8.25) is used.

**(c)** Show that the integral in Eq. (8.30) can be done analytically with the result given in Eq. (8.31). One way to do the integral is to change variables to $x \equiv \sqrt{1+y}$.

**8.5** Find the wavenumber of the mode which equals the inverse comoving Hubble radius at equality. That is, define $k_{\rm eq}$ to be equal to $a_{\rm eq}H(a_{\rm eq})$. Show that this definition

implies

$$k_{\rm eq} = \sqrt{\frac{2\Omega_{\rm m} H_0^2}{a_{\rm eq}}}. \tag{8.87}$$

Then use Eq. (2.86) to show that $k_{\rm eq}$ is given by Eq. (8.39). Show that if you define $k_{\rm eq}$ by setting it to $1/\eta_{\rm eq}$, you get a number that is somewhat lower.

**8.6** Define $a_H(k)$, the scale factor at which the wavenumber $k$ equals the comoving Hubble radius, via $a_H H(a_H) \equiv k$. Express $a_H/a_{\rm eq}$ in terms of $k$ and $k_{\rm eq}$. Show that in the limit $k \gg k_{\rm eq}$, this expression reduces to

$$\lim_{k\gg k_{\rm eq}} \frac{a_H}{a_{\rm eq}} = \frac{k_{\rm eq}}{\sqrt{2}k}. \tag{8.88}$$

**8.7** Show, using CAMB or CLASS, that the shape of the matter power spectrum is largely independent of $\Omega_{\rm m}$ when the physical density $\Omega_{\rm m}h^2$ is held fixed by changing $h$ correspondingly, and when plotting in physical units (i.e. $\mathrm{Mpc}^{-1}$). Now go back to conventional, $h$-scaled units and vary $\Omega_{\rm m}$ holding the combination $\Omega_{\rm m}h$ fixed. Explain your findings for the change in the matter power spectrum in this case.

**8.8** Solve Eq. (8.75) under the assumption that only matter, curvature, and a cosmological constant are relevant:

$$H^2(a) = H_0^2\left[\Omega_{\rm m}a^{-3} + \Omega_\Lambda + (1-\Omega_{\rm m}-\Omega_\Lambda)a^{-2}\right]. \tag{8.89}$$

**(a)** Show that $\delta_{\rm m} \propto H$ is a solution. What property of this solution makes it unsuited to describing the growth of structure in the universe?

**(b)** To obtain the second solution, try a solution of the form $u = \delta_{\rm m}/H$. Compare with Eq. (8.77).

**(c)** Now generalize Eq. (8.89) to non-constant dark energy, $\Omega_\Lambda \to \Omega_{\rm de}(a)$, with equation of state $w$. Under what conditions on $w$ does Eq. (8.77) solve Eq. (8.75)?

**8.9** Compute the growth factors in a universe with $\Omega_{\rm de} = 0.7$, $\Omega_{\rm m} = 0.3$, and $w = -0.5$. For this, solve Eq. (8.75) numerically. Plot the growth factor as a function of $a$. Compare with the cosmological constant model ($w = -1$) with the same $\Omega_{\rm de}$, $\Omega_{\rm m}$.

**8.10** Compute the change to the scale-independent growth factor due to the effect that massive neutrinos have on the expansion rate $H(a)$, via Eq. (8.75). For this, the results of Exercise 2.13 are useful. Consider two cases: a single massive species with (*i*) $m_\nu = 0.06$ eV and (*ii*) $m_\nu = 0.2$ eV.

**8.11** Compute the neutrino free-streaming scale $k_{\rm fs}$. First, determine the typical momentum of neutrinos with temperature $T_{\nu,0}/a = 1.946\,\mathrm{K}/a$. Then, calculate the typical distance $x_{\rm fs}$ a neutrino with this momentum and mass $m_\nu$ travels within a time interval $\Delta t = 1/H$. The free-streaming scale is then given by $k_{\rm fs} = 1/x_{\rm fs}$. Show that this

can be written as

$$k_{\rm fs}(a) \simeq 3.2\sqrt{a^{-2} + m_\nu^2/T_{\nu,0}^2}\, a^2 H(a)$$
$$\simeq 0.063\, h\,{\rm Mpc}^{-1}\, \frac{m_\nu}{0.1\ {\rm eV}}\, \frac{a^2 H(a)}{H_0},$$

where we have used $m_\nu/T_{\nu,0} \gg a^{-1}$ in the last equality, which applies at the late times of interest.

**8.12** Find $k_{\rm NL}$ defined in Sect. 8.1.1 for the fiducial ΛCDM model at $z = 0$, $z = 1$, and $z = 2$.

**8.13** Another popular way to characterize the amplitude of matter fluctuations on a particular scale is to compute the expected RMS overdensity in a sphere of comoving radius $R$ (not to be confused with the ratio of baryon-to-photon energy density),

$$\sigma_R^2 \equiv \langle \delta_{{\rm m},R}^2(\boldsymbol{x})\rangle. \tag{8.90}$$

Here

$$\delta_{{\rm m},R}(\boldsymbol{x}) \equiv \int d^3x' \delta_{\rm m}(\boldsymbol{x}') W_R(|\boldsymbol{x} - \boldsymbol{x}'|) \tag{8.91}$$

where $W_R(x)$ is the *tophat* window function, equal to $3/(4\pi R^3)$ for $x < R$ and 0 otherwise; the angular brackets denote the ensemble average.

**(a)** By Fourier transforming, express $\sigma_R$ in terms of an integral over the power spectrum.

**(b)** Use the transfer function from CAMB or CLASS, or your code in the case you solved Exercise 8.2, to compute $\sigma_8 \equiv \sigma_R(R = 8\,h^{-1}\,{\rm Mpc})$ for the fiducial ΛCDM cosmology.

**(c)** In the same model, plot $\sigma_R$ as a function of $R$. Since $\sigma_R$ monotonically increases toward small $R$, small scales tend to go nonlinear before large scales, the signature of a hierarchical model. Compare with Fig. 12.1.

# 9 The cosmic microwave background

The primordial perturbations set up during inflation manifest themselves in the matter distribution as well as in the radiation. By understanding the evolution of the photon perturbations, we can make predictions for the power spectrum of CMB anisotropies shown in Fig. 1.10. This evolution is again completely determined by the Einstein–Boltzmann system we derived in Chs. 5–6, and one way to go would be to code up all the relevant equations in those chapters and solve them numerically. Historically, this is a pretty good caricature of what happened. Long before we developed deep insight into the physics of anisotropies, various groups had codes that determined the expected power spectra from different models. Only much later did we come to understand why the anisotropies look like they do. In this chapter, we will develop this understanding by deriving approximate semi-analytic solutions.

Perturbations to the photons evolved completely differently before and after the epoch of recombination at $z_* \simeq 1100$. Before recombination, the photons were tightly coupled to the electrons and protons; all together they can be described as a single fluid (dubbed the "baryon–photon" fluid). After recombination, photons free-streamed from the "surface of last scattering" to us today. After an overview (Sect. 9.1) that qualitatively explains the anisotropy spectrum, Sects. 9.2–9.4 work through the physics of the baryon–photon fluid before recombination. Then, Sects. 9.5–9.6 treat the post-recombination era, culminating in the predicted spectrum of anisotropies today. Finally, Sect. 9.7 discusses how, and to what extent, the CMB power spectrum allows us to determine the cosmological parameters.

## 9.1 Overview

Let us begin as we did in the last chapter, by cheating and looking at the answers first. Fig. 9.1 shows the evolution of the perturbations to the photons. Four Fourier modes corresponding to perturbations on four different scales are shown up to the point of when they decouple from baryons at $\eta = \eta_*$. The first important point is that the photon perturbations do not grow after decoupling: since gravitational potentials in the universe are too weak to trap photons, the photons travel freely through the universe as soon as they decouple from the baryons, and essentially preserve the perturbations at the level they had at decoupling. This stands in stark contrast to the perturbations in baryons and CDM, which grow by orders of magnitudes between decoupling and today.

Before going further and examining the evolution of the different modes in more detail, a technical note: we have normalized $(\Theta_0 + \Psi)(\boldsymbol{k}, \eta)$ to the initial gravitational potential

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00015-2
Copyright © 2021 Elsevier Inc. All rights reserved.

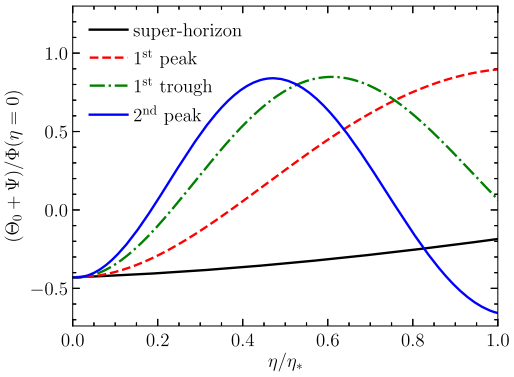



**FIGURE 9.1** Evolution of four different modes of photon perturbations before recombination at $\eta_*$, in the fiducial $\Lambda$CDM cosmology and normalized to the gravitational potential at the end of inflation. In the order of appearance in the legend (from top to bottom), the wavenumbers are $k\,[h\ \mathrm{Mpc}^{-1}] = 0.005, 0.020, 0.031, 0.039$.

$\Phi(\boldsymbol{k}, 0)$, i.e. at the end of inflation (this is directly related to the curvature perturbation $\mathcal{R}$ via Eq. (8.32)). The plot shows the sum of the gravitational potential $\Psi$ and the photon monopole $\Theta_0$ because the photons we see today had to travel out of the potentials they were in at the time of recombination. As they emerged from these potential wells, their wavelengths were stretched (if the region was overdense and $\Psi < 0$), thereby decreasing their energy; this is the gravitational redshift we encountered in Sect. 3.3.2. Thus, the temperature we see today is actually $\Theta_0$ plus $\Psi$ at recombination. Roughly speaking, then, the observed CMB anisotropies are given by an integral over $\boldsymbol{k}$ of the quantity shown in Fig. 9.1 squared, multiplied by the power spectrum of $\Phi(\boldsymbol{k}, 0)$, essentially $P_{\mathcal{R}}(k)$. So for the anisotropy power spectrum, what counts is the *amplitude* of the quantity shown in Fig. 9.1, not its sign.

The large-scale mode in Fig. 9.1 evolves hardly at all. This is generally the case for super-horizon perturbations: no causal physics can affect such perturbations with wavelengths larger than the horizon, so a super-horizon mode should exhibit little evolution. This means that when we observe the CMB anisotropies on large scales—which are determined by modes with wavelengths larger than the horizon at recombination—we are observing perturbations in their most pristine form, as they were generated at very early times, during inflation.

Fig. 9.1 shows that the smaller-scale modes evolve in a more complicated way than the super-horizon modes. Consider the curve labeled "1st peak." As the mode enters the horizon, the perturbation begins to grow until it reaches an apparent maximum at the time of recombination. If we observe anisotropies on scales corresponding to this mode, we would expect to see large fluctuations. Hence the label: the anisotropy spectrum will have a peak at the angular scales corresponding to the mode which has just reached its peak at recombination.

The mode in Fig. 9.1 that enters the horizon slightly earlier turns over so that its amplitude at recombination close to is zero. Therefore, when we observe anisotropies today

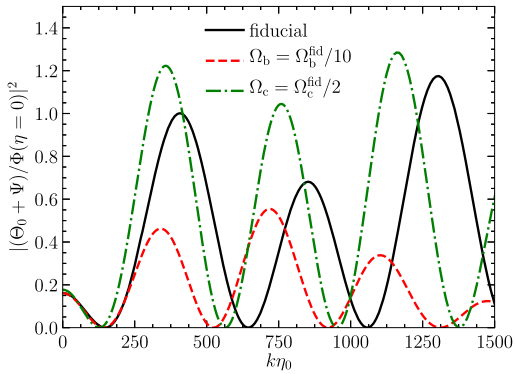



**FIGURE 9.2** Perturbations to the observed photon temperature $\Theta_0 + \Psi$ squared, normalized to the potential at the end of inflation $\Phi(k, \eta = 0)$. All curves are evaluated at $\eta = \eta_*$, where $\eta_*$ is the recombination time in the fiducial cosmology. The black solid line shows the fiducial cosmology, while the other two curves show results with a model with reduced baryons (leading to an increased oscillation frequency) and reduced CDM (leading to a suppressed asymmetry of even and odd peaks). The larger damping length $\lambda_D$ of the low-$\Omega_b$ case is clearly evident in the suppression of perturbations with $k \gtrsim 1000/\eta_0$.

corresponding to these scales, we expect very small fluctuations. There will be a trough in the anisotropy spectrum on these angular scales.

And on it goes. The curve labeled "2nd peak" entered the horizon even earlier and has gone through one full oscillation by recombination. As such, this mode has a large amplitude, and hence leads to a second peak in the anisotropy spectrum. By now the pattern of the *acoustic oscillations* becomes clear: you might expect that there will be a never-ending series of peaks and troughs in the anisotropy spectrum corresponding to modes that entered the horizon earlier and earlier. This is exactly what happens.

We can see this more clearly by looking at the spectrum of perturbations at one time, the time of recombination. Fig. 9.2 shows this spectrum for the fiducial cosmology (black solid). We do indeed see this pattern of peaks and troughs. Note that the heights of the peaks alternate: the odd peaks are higher than the even peaks in the fiducial case. To understand this feature, we can write down a cartoon version of the equation governing perturbations:

$$\Theta_0'' + k^2 c_s^2 \Theta_0 = F \tag{9.1}$$

where $F$ is a driving force due to gravity and $c_s$ is the sound speed of the combined baryon–photon fluid (we will derive it below, Eq. (9.21)). This is the equation of a forced harmonic oscillator (see Box 9.1). Qualitatively, it predicts the oscillations we have seen above.

First, the oscillation frequency is determined by the ratio of the spring constant and the mass; in case of the baryon–photon fluid this means that the oscillation frequency becomes larger if we decrease the mass loading of the fluid, i.e. $\Omega_b$. That is, the fewer baryons there are, the faster is the speed of sound propagation. This can be seen in the low-$\Omega_b$ curve in Fig. 9.2.

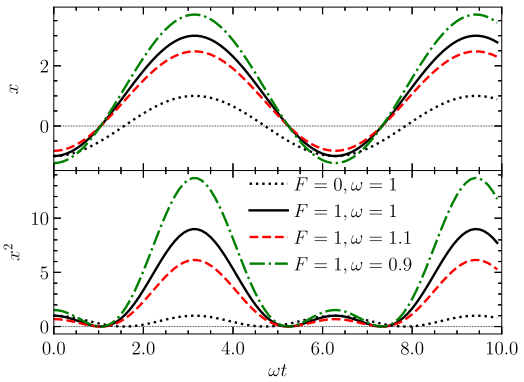



**FIGURE 9.3** The forced harmonic oscillator solution discussed in Box 9.1.

Second, the external force (or more precisely, acceleration) $F$ sets the asymmetry in the odd and even peaks: the larger $F$, and the lower the frequency, the larger the asymmetry becomes. Consider an initial overdensity, and the sign of $F$ is such that it tends to increase the overdensity. When the baryon–photon fluid begins to contract, its self-gravity and the external force act in consort, leading to a stronger contraction than would have been the case for $F = 0$. Conversely, as the pressure wins and pushes the plasma outward to maximum expansion, it acts against the external force and thus leads to an underdensity with smaller amplitude than it would achieve if $F = 0$. In the case of the baryon–photon fluid before decoupling, the external acceleration is supplied by the gravitational potentials sourced by dark matter. Hence, the asymmetry between even and odd peaks is a direct probe of the amount of dark matter $\Omega_c$, which is visible as the reduced asymmetry in the low-$\Omega_c$ case in Fig. 9.2.

This reasoning has qualitative merit, but clearly misses some of the details. For example, the asymmetry between even and odd peaks appears to be of opposite sign for the low-baryon case. This feature is a combination of several physical effects, including increased damping (which we will discuss next) and the time dependence of the actual force term (which we have assumed to be constant for this discussion). This is fine—we will treat all these effects precisely in Sect. 9.3.

## 9.1 The forced harmonic oscillator

Consider a simple harmonic oscillator with mass $m$ and force constant $K$. In addition to the restoring force, the oscillator is acted on by an external force $F$ (really acceleration). Thus the full force is $mF - Kx$ where $x$ is the oscillator's position. The equation of motion is

$$\ddot{x} + \frac{K}{m}x = F. \tag{9.2}$$

The term on the right-hand side—representing the external force—is driving the oscillator to large values of $x$ (assuming that $F > 0$). The restoring force on the other hand tries to keep the oscillator as close to the origin as possible. The result is that oscillations occur around a different zero point, at positive $x$.

The solution to Eq. (9.2) is the sum of the general solution to the homogeneous equation (with the right-hand side set to zero) and a particular solution. The general solution has two modes, best expressed as a sine and cosine with arguments $\omega t$, with an angular frequency given by $\omega \equiv \sqrt{K/m}$. A particular solution to Eq. (9.2) is a constant $x = F/\omega^2$, so the full solution is the sum of the sine and cosine modes plus this constant. Let us assume that the oscillator is initially at rest. Then, since $\dot{x}(0)$ is proportional to the coefficient of the sine mode, this coefficient must vanish, leaving

$$x = A\cos(\omega t) + \frac{F}{\omega^2}. \tag{9.3}$$

This solution is shown in the upper panel of Fig. 9.3. The dotted line is the unforced solution: oscillations about the origin. The solid curve shows the forced solutions with the same frequency: the oscillations are not around $x = 0$ as they would be if the system was unforced. Once an external force is introduced, the zero point of the oscillations shifts in the direction of the force. The dashed and dot-dashed curves illustrate the effect of varying the frequency while the force is kept fixed. The zero-point shift is more dramatic for lower frequencies. The bottom panel shows the square of the oscillator position as a function of time, which is analogous to Fig. 9.2. All three cases show a series of peaks at $t = n\pi/\omega$ corresponding to the minima/maxima of the cosine mode. (Note that if only the sine mode was present, these peaks would be at $t = (2n+1)\pi/\omega$; in general, if both cosine and sine modes are present, peaks can appear anywhere in time.) The heights of all peaks are identical in the case of the unforced oscillator. In the forced case, though, the height of the odd peaks at $t = \pi/\omega, 3\pi/\omega, \ldots$ is greater than that of the even peaks at $t = 0, 2\pi/\omega, \ldots$.

To summarize, the behavior of the forced oscillator is determined by two parameters: the reduced spring constant $K/m$ which sets the oscillation frequency, and the external force $F$ which sets the zero point and the asymmetry between even and odd peaks.

Beyond the oscillations, the damping that is visible on small scales $k\eta_0 \gtrsim 500$ for the low-$\Omega_b$ case in Fig. 9.2 is a further important effect in the physics of radiation perturbations (in the fiducial model, the damping sets in at higher values of $k$). To understand this, we need to remember that the approximation of the photons and baryons moving together as a single fluid is valid only if the scattering rate of photons off of electrons is infinite. In reality this condition is not met: photons travel a finite distance in between scatters.

Consider the path of a single photon as it scatters off a sea of electrons (Fig. 9.4). It travels a mean comoving distance $\lambda_{\rm MFP}$ in between each scatter. In our case this distance is $(n_e\sigma_{\rm T}a)^{-1} = -1/\tau'$, where $\tau$ is the optical depth defined in Eq. (5.33). If the density $n_e$ of electrons is very large, then the mean free path is correspondingly small. Over the course of a Hubble time, $H^{-1}$, a photon scatters of order $n_e\sigma_{\rm T}H^{-1}$ times (simply the product of the rate and the time), performing a random walk. We know that the total distance traveled in the course of a random walk is the mean free path times the square root of the total number of steps. Therefore, a cosmological photon moves a mean comoving distance

$$\begin{aligned} \lambda_D &\sim \lambda_{\rm MFP}\sqrt{n_e\sigma_{\rm T}H^{-1}} \\ &= \frac{1}{\sqrt{n_e\sigma_{\rm T}H}}\frac{1}{a} \end{aligned} \tag{9.4}$$

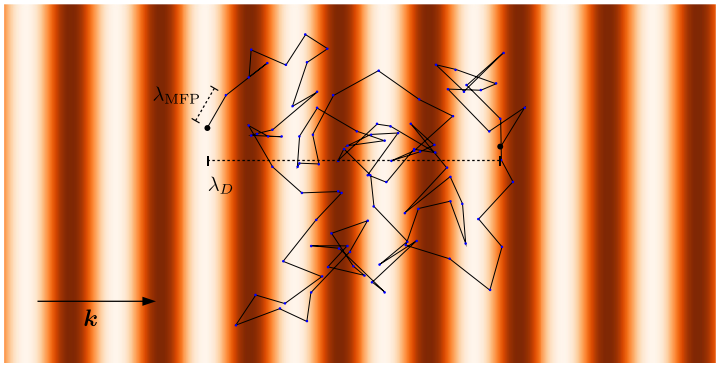



**FIGURE 9.4** Photon diffusion through the electron gas. The line illustrates the random walk of a photon. Scattering events are denoted as points, while the large dots denote the initial and final locations of the photon. Each scattering event is separated by a distance of order the mean free path $\lambda_{\rm MFP}$, while the photon has overall moved a distance of order the damping length $\lambda_D$ between its initial and final positions. Perturbations with $k \gtrsim 1/\lambda_D$, like the one sketched here, will be washed out by the diffusion.

in a Hubble time. Any perturbation on scales smaller than $\lambda_D$ can be expected to be washed out, because the many photons diffusing over a region of order $\lambda_D$ will have restored the region to a single mean temperature (as you can glean from Fig. 9.4). In Fourier space this effect corresponds to damping of all high-$k$ modes. This crude estimate predicts an $\Omega_{\rm b}$ dependence consistent with Fig. 9.2: reducing the baryon density leads to a larger $\lambda_D$ (since $n_e$ is proportional to $\Omega_{\rm b}$ when the universe is ionized), and hence stronger damping.

The final step is to relate the perturbations at recombination, as depicted in Fig. 9.2, to the anisotropies we observe today. The math of this is a little complicated, but the physics is perfectly straightforward. Consider one Fourier mode, i.e. a plane-wave perturbation. Fig. 9.5 shows the temperature variations for one mode at recombination. Photons from hot and cold spots separated by a typical comoving distance $k^{-1}$ travel to us coming from an angular separation $\theta \simeq k^{-1}/\chi_*$ where $\chi_* = \eta_0 - \eta_*$ is the comoving distance between us and the surface of last scattering.[1] If we decompose the temperature field into multipole moments, then an angular scale $\theta$ roughly corresponds to $1/l$. So, using the fact that $\eta_* \ll \eta_0$, we project inhomogeneities on scales $k$ onto anisotropies on angular scales $l \simeq k\eta_0$.

There is one final caveat to this picture of free-streaming. We have been implicitly assuming that nothing happens to the photons on their journey from the last-scattering surface to Earth. This is not completely true. While gravitational potentials are constant deep in matter domination, they do evolve right after recombination (due to the presence of radiation) and at late times (due to dark energy). Evolving potentials produce additional perturbations to the photons via the *integrated Sachs–Wolfe (ISW) effect*. Finally, the universe is no longer completely neutral at redshifts $z \lesssim 10$, and the presence of free elec-

[1] This is true only in a Euclidean universe. In an open universe, for example, the angular diameter distance to the last-scattering surface is larger, so the same physical scale is projected onto a smaller angular scale (see Fig. 9.14).

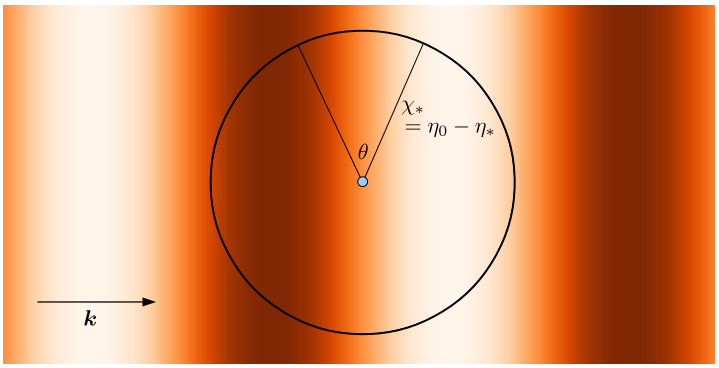



**FIGURE 9.5** Perturbations in the temperature due to a plane wave with wavenumber $k$. Hot and cold regions are shaded light and dark. After recombination, photons from the hot and cold spots travel freely to us, denoted by the blue dot at the center. This $\boldsymbol{k}$-mode contributes anisotropy on a scale $\theta \sim k^{-1}/\chi_*$, where $\chi_* = \eta_0 - \eta_*$ is the comoving distance to the last-scattering surface.

trons leads to scattering of CMB photons which in turn slightly dampens the anisotropies. This is, in a nutshell, how primordial perturbations are processed to form the present-day anisotropy spectrum.[2] Now let us work through each step again quantitatively.

## 9.2 Large-scale anisotropies

To find the large-scale solution for the photon perturbation, we make use of the super-horizon equation (8.16). This immediately tells us that $\Theta_0 = -\Phi$ plus a constant. The initial conditions, specifically Eq. (7.91), are such that $\Theta_0(\eta = 0) = \Phi(\eta = 0)/2$, so the constant is $\mathcal{R}$, where $\mathcal{R}$ is the conserved curvature perturbation set during inflation. We have an exact expression for the large-scale evolution of $\Phi$, Eq. (8.31). Since recombination takes place long after the epoch of equality, we can take the $y \gg 1$ limit of this expression and Eq. (8.32) holds: $\Phi = (3/5)\mathcal{R}$. Therefore, at recombination, large-scale photon perturbations satisfy

$$\begin{aligned}\Theta_0(\boldsymbol{k}, \eta_*) &= -\Phi(\boldsymbol{k}, \eta_*) + \mathcal{R}(\boldsymbol{k}) \\ &= \frac{2}{5}\mathcal{R}(\boldsymbol{k}) = \frac{2}{3}\Phi(\boldsymbol{k}, \eta_*).\end{aligned} \tag{9.5}$$

The observed anisotropy is $\Theta_0 + \Psi$, which to a good approximation is $\Theta_0 - \Phi$ since $\Psi \simeq -\Phi$. Therefore,

$$(\Theta_0 + \Psi)(\boldsymbol{k}, \eta_*) = -\frac{1}{3}\Phi(\boldsymbol{k}, \eta_*) = -\frac{1}{5}\mathcal{R}(\boldsymbol{k}). \tag{9.6}$$

Eq. (9.6) will be useful to us when we compute the large-scale anisotropy spectrum.

[2] We will learn about one more, *nonlinear* effect on the CMB photons in Sect. 13.3: their deflection by gravitational potential wells.

Another useful way of expressing the large-scale perturbations at recombination is in terms of the dark-matter density field. The initial conditions derived in Ch. 7 were that $\delta_c = \mathcal{R}$. Integrating the large-scale evolution equation, $\delta_c{}' = -3\Phi'$ (Eq. (8.17)), leads to

$$\delta_c(\boldsymbol{k}, \eta_*) = \mathcal{R}(\boldsymbol{k}) - 3\left[\Phi(\boldsymbol{k}, \eta_*) - \frac{2}{3}\mathcal{R}(\boldsymbol{k})\right], \tag{9.7}$$

where the factor in brackets reduces to zero if we set $\eta_* \to 0$, enforcing the correct initial conditions. Thus,

$$\delta_c(\boldsymbol{k}, \eta_*) = \frac{6}{5}\mathcal{R}(\boldsymbol{k}) = 2\Phi(\boldsymbol{k}, \eta_*). \tag{9.8}$$

So the observed anisotropy expressed in terms of the dark matter overdensity is

$$(\Theta_0 + \Psi)(\boldsymbol{k}, \eta_*) = -\frac{1}{6}\delta_c(\boldsymbol{k}, \eta_*). \tag{9.9}$$

This relation contains an interesting piece of information. We see that the observed temperature perturbation of an overdense region is, surprisingly, *negative*. Large-scale overdense regions do indeed contain hotter photons at recombination than do underdense regions: i.e., $\Theta_0 > 0$ when $\Psi < 0$. However, to get to us today, these photons must travel out of their potential wells. In so doing they lose energy, and this energy loss more than compensates for the fact that the photons were initially hotter than average: i.e., $\Theta_0 + \Psi$ is negative when $\Psi < 0$. To sum up, when we observe large-scale hot spots on the CMB sky today, we are actually observing regions that were underdense at the time of recombination.

The other important feature of Eq. (9.9) is the coefficient 1/6. It enables us to relate "$\delta T/T$" (the left-hand side) to "$\delta\rho/\rho$" (the right). Roughly, an anisotropy of order $10^{-5}$ corresponds to an overdensity of $6 \times 10^{-5}$. As discussed at the end of the previous chapter, one of the important questions that must be addressed by any viable cosmological model is whether the observed CMB anisotropy is consistent with the matter-density perturbations needed to form the observed structure by today. This factor of 6 is essential to accomplish this.

## 9.3 Acoustic oscillations

Before electrons and nuclei began forming atoms, so before $\eta_*$, the mean free path for a photon was much smaller than the horizon. Compton scattering caused the electron–proton fluid to be tightly coupled with the photons. We now proceed to explore this regime quantitatively using the Boltzmann equations.

### 9.3.1 Tightly-coupled limit of the Boltzmann equations

The tightly-coupled limit applies when the mean free path of the photons is much smaller than the scales of interest. Essentially, this is equivalent to $\tau \gg 1$. We want to argue that

in the $\tau \gg 1$ limit, the only nonnegligible moments, $\Theta_l$, are the monopole ($l = 0$) and the dipole ($l = 1$). All others are suppressed. In this sense, photons behave just like a fluid, which can be described with only two variables: the density $\rho$ and the (longitudinal) velocity $u$. In order to show this, let us go back to the Boltzmann equation (5.67) for photons. We want to turn this differential equation for $\Theta(k, \eta, \mu)$ into an infinite set of coupled equations for $\Theta_l(k, \eta)$. The strategy is to multiply by $\mathcal{P}_l(\mu)$ and then integrate over $\mu$. Using Eq. (5.66), the Boltzmann equation for $l > 2$ becomes

$$\Theta_l' + \frac{k}{(-i)^{l+1}} \int_{-1}^{1} \frac{d\mu}{2} \mu \mathcal{P}_l(\mu) \Theta(\mu) = \tau' \Theta_l \qquad (l > 2). \tag{9.10}$$

Note that all other terms in the Boltzmann equation (e.g., $-\Phi'$) scale either as $\mu^0$, $\mu^1$, or $\mu^2$, so they disappear after the integral over $\mu$ against $\mathcal{P}_l$ with $l > 2$. To do the integral in the second term here, we make use of the recurrence relation for Legendre polynomials, Eq. (C.3), to get

$$\Theta_l' - \frac{kl}{2l+1} \Theta_{l-1} + \frac{k(l+1)}{2l+1} \Theta_{l+1} = \tau' \Theta_l. \tag{9.11}$$

Let us consider the order of magnitude of the terms in Eq. (9.11). The first term on the left is of order $\Theta_l/\eta$ which is much smaller than the term on the right which is enhanced by the factor $\tau'$. Neglecting the $\Theta_{l+1}$ term for the moment, this tells us that in the tightly-coupled regime

$$\Theta_l \sim -\frac{k}{\tau'} \frac{l}{2l+1} \Theta_{l-1}. \tag{9.12}$$

Recalling that the mean free path is $\lambda_{\rm MFP} = -1/\tau'$, the prefactor is $k\lambda_{\rm MFP}$. Thus, for all modes with wavelengths much larger than the mean free path, we have $\Theta_l \ll \Theta_{l-1}$. This also furnishes our justification for throwing out the $\Theta_{l+1}$ term in making our estimate: a similar relation holds between $\Theta_{l+1}$ and $\Theta_l$. It is easy to verify that if we neglect the multipole moments of polarization, $\Theta_2$ is likewise suppressed. We will see in the next chapter that the suppression also applies to polarization. To summarize then, all moments with $l > 1$ are very small compared to the monopole and dipole.

Before making use of this fact and deriving the tightly-coupled equations in the limit in which only the monopole and dipole are nonzero (the fluid approximation), we want to understand *why* higher moments are damped in a tightly-coupled environment. Indeed this observation is extremely important not only in cosmology but in all settings in which the fluid approximation is used. Consider a plane-wave perturbation as depicted in Fig. 9.6. An observer sitting at the center of the perturbation sees photons arriving from a distance of order the mean free path, $-1/\tau'$. Very little anisotropy is then induced by a large-scale perturbation, one with $k/|\tau'| \ll 1$ as shown in Fig. 9.6, since the temperature hardly varies within the region from which the observed photons originate. This does not hold for perturbations with very small wavelengths (with $k/|\tau'| \sim 1$). In fact, though, those

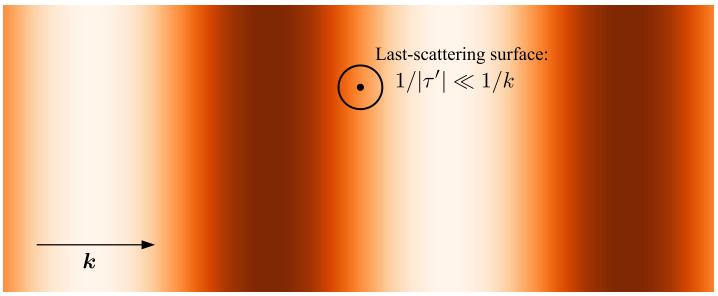



**FIGURE 9.6** Anisotropies in the tightly-coupled era, for a perturbation that is of much larger scale than the mean free path of the photons $1/|\tau'|$. The photons measured by an observer (denoted by the dot) come from within a distance $1/|\tau'|$ away that is much smaller than the wavelength of the mode. Hence, an observer sees photons arriving from all angles with virtually identical temperatures; more precisely, she will measure a monopole and a small dipole, with all higher moments being negligible.

modes are strongly damped by photon diffusion, since their wavelengths are much smaller than the damping scale.

Armed with this knowledge, we can now turn to the equations for the first two moments, which—after disposing of $\Theta_2$—read

$$\Theta_0' + k\Theta_1 = -\Phi', \tag{9.13}$$

$$\Theta_1' - \frac{k\Theta_0}{3} = \frac{k\Psi}{3} + \tau'\left[\Theta_1 - \frac{iu_\mathrm{b}}{3}\right]. \tag{9.14}$$

These follow by multiplying Eq. (5.67) by $\mathcal{P}_0(\mu)$ and $\mathcal{P}_1(\mu)$ and integrating over $\mu$. They are supplemented by the equations for the electron–baryon fluid, Eqs. (5.71) and (5.72). Let us first rewrite the velocity equation, (5.72), as

$$u_\mathrm{b} = -3i\Theta_1 + \frac{R}{\tau'}\left[{u_\mathrm{b}}' + \frac{a'}{a}u_\mathrm{b} + ik\Psi\right], \tag{9.15}$$

where, recall, the baryon-to-photon energy ratio $R = R(\eta)$ is defined as

$$R \equiv \frac{3\rho_\mathrm{b}}{4\rho_\gamma}. \tag{9.16}$$

The second term on the right-hand side of Eq. (9.15) is much smaller than the first since it is suppressed by a relative factor of order $1/\tau'\eta$ and $k/\tau'$, respectively. Thus, to lowest order, $u_\mathrm{b} = -3i\Theta_1$. A systematic way to expand, then, is to use this lowest-order expression everywhere in the second term, leading to

$$u_\mathrm{b} \simeq -3i\Theta_1 + \frac{R}{\tau'}\left[-3i\Theta_1' - 3i\frac{a'}{a}\Theta_1 + ik\Psi\right]. \tag{9.17}$$

Now let us insert this expression into Eq. (9.14), eliminating $u_{\rm b}$. After rearranging terms, we find

$$\Theta_1' + \frac{a'}{a}\frac{R}{1+R}\Theta_1 - \frac{1}{3}\frac{k}{1+R}\Theta_0 = \frac{k\Psi}{3}. \tag{9.18}$$

We now have two first-order coupled equations for the first two photon moments, Eqs. (9.13) and (9.18). We can turn these into one second-order equation by differentiating Eq. (9.13) and using Eq. (9.18) to eliminate $\Theta_1'$:

$$\Theta_0'' + \frac{k^2}{3}\Psi - \frac{a'}{a}\frac{R}{1+R}k\Theta_1 + \frac{1}{3}\frac{k^2}{1+R}\Theta_0 = -\Phi''. \tag{9.19}$$

Finally, we use Eq. (9.13) to eliminate $\Theta_1$ here. This leaves

$$\begin{aligned}\Theta_0'' + \frac{a'}{a}\frac{R}{1+R}\Theta_0' + k^2c_s^2\Theta_0 &= F(k,\eta),\\ F(k,\eta) &\equiv -\frac{k^2}{3}\Psi - \frac{a'}{a}\frac{R}{1+R}\Phi' - \Phi'',\end{aligned} \tag{9.20}$$

where we have defined the force function on the right as $F$ and the sound speed of the fluid as

$$c_s(\eta) \equiv \sqrt{\frac{1}{3(1+R[\eta])}}. \tag{9.21}$$

The sound speed depends on the baryon density in the universe. If the baryon density is negligible compared to that of the radiation, $c_s$ has the standard value for a relativistic fluid, $c_s = 1/\sqrt{3}$. The presence of baryons makes the fluid heavier, thereby lowering the sound speed; this is analogous to the inverse mass in the term $(K/m)x$ in Eq. (9.2) of the forced harmonic oscillator. We will see shortly that the fluid oscillates in both space and time, with a period determined by the sound speed, and hence by the baryon density. Note that Eq. (9.20) is the "grown-up" version of Eq. (9.1); it differs only through the $\Theta_0'$ drag term (see Exercise 9.2), and the correct, time-dependent force term. The presence of the drag term does not change any of the qualitative conclusions we reached in Sect. 9.1. Finally, note that $\Phi$ enters on the right in a very similar way as $\Theta_0$ does on the left. An alternate version of Eq. (9.20) takes advantage of this:

$$\left\{\frac{d^2}{d\eta^2} + \frac{R'}{1+R}\frac{d}{d\eta} + k^2c_s^2\right\}[\Theta_0+\Phi](\boldsymbol{k},\eta) = \frac{k^2}{3}\left[\frac{1}{1+R}\Phi - \Psi\right](\boldsymbol{k},\eta). \tag{9.22}$$

Notice that the combination $\Theta_0 + \Phi$, which is convenient in the context of tight coupling, is *not* the combination $\Theta_0 + \Psi \simeq \Theta_0 - \Phi$ which yields the observed CMB temperature.

### 9.3.2 Tightly-coupled solutions

The equation we have derived governing acoustic oscillations of the photon-baryon fluid, (9.22), is a second-order ordinary differential equation. To solve it, we will again (as in Sect. 8.3.1) use Green's method to find the full solution. First we find the two solutions to the homogeneous equation. Then we use these to construct the particular solution.

In principle, to obtain the homogeneous solutions, we must solve the damped, harmonic oscillator equation (9.22) with the right-hand side equal to zero. In practice, the drag term is of order $R(\Theta_0+\Phi)/\eta^2$ while the pressure term is much larger, of order $k^2c_s^2(\Theta_0+\Phi)$ (more precisely, it is larger when modes are within the horizon or when $R$ is small). Physically, the time scale of the pressure-induced oscillations is much smaller than the expansion time over which the drag operates. To a first approximation, then, let us neglect the drag term and simply obtain the oscillating solutions; you can rectify this by applying the WKB approximation in Exercise 9.5. In this limit, the two homogeneous solutions are

$$S_1(k,\eta)=\sin[kr_s(\eta)];\qquad S_2(k,\eta)=\cos[kr_s(\eta)] \tag{9.23}$$

where we have defined the *sound horizon* as

$$r_s(\eta)\equiv\int_0^\eta d\tilde{\eta}c_s(\tilde{\eta}). \tag{9.24}$$

Since $c_s$ is the sound speed, the sound horizon is the comoving distance traveled by a sound wave by time $\eta$.

The tightly-coupled solution for the photon temperature can be constructed from the homogeneous solutions of Eq. (9.23):

$$\begin{aligned}\Theta_0(\boldsymbol{k},\eta)+\Phi(\boldsymbol{k},\eta)&=C_1(\boldsymbol{k})S_1(\eta)+C_2(\boldsymbol{k})S_2(\eta)\\&\quad+\frac{k^2}{3}\int_0^\eta d\tilde{\eta}[\Phi(\boldsymbol{k},\tilde{\eta})-\Psi(\boldsymbol{k},\tilde{\eta})]\frac{S_1(\tilde{\eta})S_2(\eta)-S_1(\eta)S_2(\tilde{\eta})}{S_1(\tilde{\eta})S_2'(\tilde{\eta})-S_1'(\tilde{\eta})S_2(\eta)}.\end{aligned} \tag{9.25}$$

Here again, we have dropped all occurrences of $R$ except in the arguments of the rapidly varying sines and cosines. That is, the argument of $S_1$, for example, is still taken to be $kr_s$ with its nonzero value of $R$. We fix the constants $C_1$ and $C_2$ in Eq. (9.25) by matching to the initial conditions, when both $\Theta_0$ and $\Phi$ are constants. Since $\Theta_0'$ and $\Phi'$ vanish initially, the coefficient of the sine term $C_1$ must vanish. Then we have $C_2(\boldsymbol{k})=\Theta_0(\boldsymbol{k},0)+\Phi(\boldsymbol{k},0)$. The denominator in the integrand reduces to $-kc_s(\tilde{\eta})\to-k/\sqrt{3}$ in the limit in which we are working. Finally, the difference of the products in the numerator of the integrand is simply $-\sin[k(r_s-r_s')]$, so

$$\begin{aligned}\Theta_0(\boldsymbol{k},\eta)+\Phi(\boldsymbol{k},\eta)&=[\Theta_0(\boldsymbol{k},0)+\Phi(\boldsymbol{k},0)]\cos(kr_s)\\&\quad+\frac{k}{\sqrt{3}}\int_0^\eta d\tilde{\eta}[\Phi(\boldsymbol{k},\tilde{\eta})-\Psi(\boldsymbol{k},\tilde{\eta})]\sin[k(r_s(\eta)-r_s(\tilde{\eta}))].\end{aligned} \tag{9.26}$$

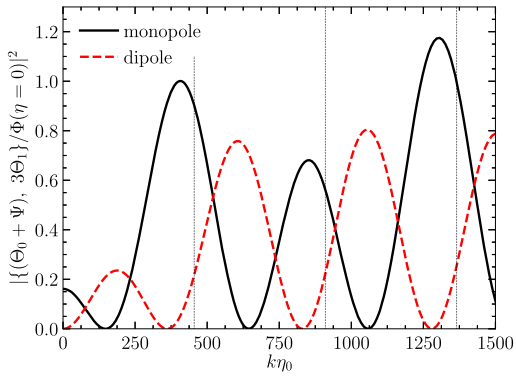



**FIGURE 9.7** The monopole $\Theta_0 + \Psi$ and dipole $3\Theta_1$ at recombination in the fiducal concordance cosmology. The dashed vertical lines indicate the rough analytic peak locations of Eq. (9.27). The dipole is completely out of phase with the monopole, and vanishes for the longest-wavelength modes that have not entered the horizon by recombination.

Eq. (9.26) is an expression for the anisotropy in the tightly-coupled limit, first derived by Hu and Sugiyama (1995). It shows the characteristic features of the initial conditions from inflation, which only excite the cosine mode. This follows directly from the fact that perturbations were generated when they exited the horizon during inflation, and remained constant outside the horizon. The pure-cosine initial conditions lead to the coherent oscillations in $\Theta + \Phi$. Alternative scenarios, which generate the perturbations when the modes enter the horizon, would typically predict that both sine and cosine modes are present. In that case, there would not be a clear peak/trough structure in $\Theta_0$.

In fact, Eq. (9.26) accurately predicts the peak positions obtained from a full numerical solution, i.e. the extrema of $\Theta_0 + \Phi$ at $\eta = \eta_*$ as a function of $k$ as in Fig. 9.2. In order to evaluate Eq. (9.26) precisely, we need to perform the numerical integral over $(\Phi - \Psi)(\boldsymbol{k}, \tilde{\eta})$ in the second term. You can do this in Exercise 9.6. Here, let us simplify further. In the limit that the first term in Eq. (9.26) dominates, the peaks should appear at the extrema of $\cos(kr_s)$, i.e., at

$$k_{\rm pk} = n\pi/r_s \qquad n = 1, 2, \ldots \tag{9.27}$$

These rough peak positions are indicated in Fig. 9.7. The predicted $k_{\rm pk}$ is within 10% of the full numerical result.

In addition to the monopole, the photon distribution has a nonnegligible dipole at recombination. Using Eq. (9.13), we can obtain an analytic solution for the dipole by differentiating Eq. (9.26):

$$\Theta_1(\boldsymbol{k}, \eta) = \frac{1}{\sqrt{3}}[\Theta_0(\boldsymbol{k}, 0) + \Phi(\boldsymbol{k}, 0)]\sin(kr_s) - \frac{k}{3}\int_0^\eta d\tilde{\eta}[\Phi(\boldsymbol{k}, \tilde{\eta}) - \Psi(\boldsymbol{k}, \tilde{\eta})]\cos[k(r_s(\eta) - r_s(\tilde{\eta}))]. \tag{9.28}$$

The first term is completely out of phase with the monopole ($\sin(kr_s)$ versus $\cos(kr_s)$). Fig. 9.7 shows that this feature remains even after accounting for the integral term. This mismatch of phase will have important implications for the final anisotropy spectrum.

## 9.4 Diffusion damping

The second ingredient we need in order to accurately describe CMB spectra is diffusion damping. To analyze diffusion quantitatively, we must return to the equations for the moments of the photon distribution, Eqs. (9.11), (9.13), and (9.14). Until now, we have neglected $\Theta_2$ and all higher moments. Diffusion is characterized by a small but nonnegligible quadrupole.

We must therefore supplement the set of equations we wrote down in the last section with an equation for the quadrupole, $\Theta_2$. Our task is somewhat simplified by the fact that we will be interested in phenomena occurring only on small scales. On these scales, the gravitational potentials $\Phi$, $\Psi$ are much smaller than the radiation perturbations, specifically they are smaller by a factor $(aH/k)^2$ (see for example Eq. (6.80); we have used this result already when treating small scales in Ch. 8). Our tightly-coupled hierarchy of moments, i.e. that each successive moment is suppressed by a higher power of $1/\tau'$, continues to hold. Thus we will need to keep only the $l = 2$ moment; all higher ones can be neglected. With these approximations, we have

$$\Theta_0' + k\Theta_1 = 0, \tag{9.29}$$

$$\Theta_1' + k\left(\frac{2}{3}\Theta_2 - \frac{1}{3}\Theta_0\right) = \tau'\left(\Theta_1 - \frac{iu_\mathrm{b}}{3}\right), \tag{9.30}$$

$$\Theta_2' - \frac{2k}{5}\Theta_1 = \frac{9}{10}\tau'\Theta_2. \tag{9.31}$$

Here, we have again neglected polarization. These three equations need to be supplemented with an equation for $u_\mathrm{b}$. This is best expressed as a slight rewriting of Eq. (9.15):

$$3i\Theta_1 + u_\mathrm{b} = \frac{R}{\tau'}\left[{u_\mathrm{b}}' + \frac{a'}{a}u_\mathrm{b}\right], \tag{9.32}$$

where again we have dropped the gravitational potential.

Let us write the time dependence of the velocity as

$$u_\mathrm{b} \propto e^{i\int \omega d\tilde{\eta}} \tag{9.33}$$

and similarly for all other variables. We already know that $\omega \simeq kc_s$ in the tightly-coupled limit. Now we are searching for damping, an imaginary part of $\omega$. Since damping occurs on small scales, we have $k \gg 1/\eta \sim a'/a$, which implies that also the real part of the frequency satisfies $\omega \gg a'/a$. Then,

$$|{u_\mathrm{b}}'| = |i\omega u_\mathrm{b}| \gg \frac{a'}{a}|u_\mathrm{b}|. \tag{9.34}$$

So we can drop the second term on the right in Eq. (9.32) and the velocity equation becomes

$$\begin{aligned} u_{\rm b} &= -3i\Theta_1\left[1-\frac{i\omega R}{\tau'}\right]^{-1} \\ &\simeq -3i\Theta_1\left[1+\frac{i\omega R}{\tau'}-\left(\frac{\omega R}{\tau'}\right)^2\right] \end{aligned} \tag{9.35}$$

where we have expanded out to $\tau'^{-2}$ because $u_{\rm b}+3i\Theta_1$ is multiplied by $\tau'$ in Eq. (9.30).

The equation for the second moment of the photon field, (9.31), can be reduced similarly. First we can drop the $\Theta_2'$ term since it is much smaller than $\tau'\Theta_2$. This leaves simply

$$\Theta_2 = -\frac{4k}{9\tau'}\Theta_1, \tag{9.36}$$

which shows that our approximation scheme is controlled: higher moments are suppressed by additional powers of $k/\tau'$. The equation for the zeroth moment becomes

$$i\omega\Theta_0 = -k\Theta_1. \tag{9.37}$$

Inserting all of these into Eq. (9.30) gives the dispersion relation

$$i\omega - \frac{8k^2}{27\tau'} + \frac{k^2}{3i\omega} = \tau'\left(1-\left[1+\frac{i\omega R}{\tau'}-\left(\frac{\omega R}{\tau'}\right)^2\right]\right). \tag{9.38}$$

Collecting terms we get

$$\omega^2(1+R) - \frac{k^2}{3} + \frac{i\omega}{\tau'}\left[\omega^2R^2+\frac{8k^2}{27}\right] = 0. \tag{9.39}$$

The first two terms on the left, the leading ones in the expansion of $1/\tau'$, recover the result of the previous section, that the frequency is the wavenumber times the speed of sound (there is no forcing term now, since we have neglected the potentials on small scales). We can write the frequency as this zeroth-order piece plus a first-order correction, $\delta\omega$. Then, inserting the zeroth-order part into the terms inversely proportional to $\tau'$ leads to

$$\delta\omega = -\frac{ik^2}{2(1+R)\tau'}\left[c_s^2R^2+\frac{8}{27}\right]. \tag{9.40}$$

Therefore, the time dependence of the perturbations is

$$\Theta_0, \Theta_1 \sim \exp\left\{ik\int d\tilde{\eta}c_s(\tilde{\eta})\right\}\exp\left\{-\frac{k^2}{k_D^2}\right\} \tag{9.41}$$

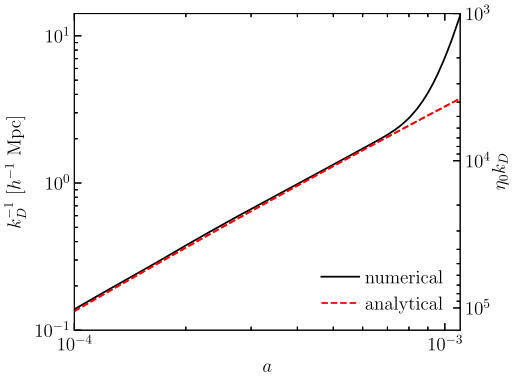



**FIGURE 9.8** Damping scale as a function of the scale factor. The solid line is obtained from numerically integrating over the standard recombination history, while the dashed curve uses the approximation of Eq. (9.44) which assumes electrons remain ionized. The right axis shows the equivalent $k_D\eta_0$; damping occurs on angular scales $l > k_D\eta_0$.

where the damping wavenumber is defined via

$$\frac{1}{k_D^2(\eta)} \equiv \int_0^\eta \frac{d\tilde{\eta}}{6(1+R)n_e\sigma_T a(\tilde{\eta})}\left[\frac{R^2}{1+R}+\frac{8}{9}\right]. \tag{9.42}$$

Putting aside factors of order unity, this equation says that $\lambda_D \sim 1/k_D \sim [\eta/n_e\sigma_T a]^{1/2}$, which agrees with our heuristic estimate at the beginning of this chapter (since $\eta \simeq 1/aH$).

As a first estimate of the damping scale, we can work in the prerecombination regime, in which all electrons (except those in helium) are free. In Ch. 4 we estimated the optical depth in this limit, but ignored helium. The mass fraction of helium is denoted as $Y_P$ and is approximately 0.24. Since each helium nucleus contains four nucleons, the ratio of helium to the total number of nuclei is $Y_P/4$. Each of these absorbs two electrons (one for each proton), so when counting the number of free electrons before hydrogen recombination, we must multiply our estimate of Eq. (4.43) by $1 - Y_P/2$. Using the fact that $H_0 = 3.33 \times 10^{-4}\,h$ Mpc$^{-1}$, we have, long before recombination,

$$n_e\sigma_T a = 2.3\times10^{-5}\,\mathrm{Mpc}^{-1}(\Omega_b h^2)a^{-2}\left(1-\frac{Y_P}{2}\right). \tag{9.43}$$

Using this, you can show (Exercise 9.8) that an approximation for the damping scale is

$$k_D^{-2} = 3.1\times10^6\,\mathrm{Mpc}^2 a^{5/2} f_D(a/a_{eq})\left(\Omega_b h^2\right)^{-1}\left(1-\frac{Y_P}{2}\right)^{-1}\left(\Omega_m h^2\right)^{-1/2} \tag{9.44}$$

where $f_D$, defined in Eq. (9.89), goes to 1 as $a/a_{eq}$ gets large.

Fig. 9.8 shows the evolution of the damping scale before recombination. Neglecting recombination is a good approximation at early times but, as expected, leads to quantitative errors near $\eta_*$ when Eq. (9.43) is used, since the free electron density does not accurately

account for the electrons swept up into neutral hydrogen. When ignoring recombination, $k_D$ scales as $\Omega_{\rm b}^{1/2}$; the details of recombination change this simple scaling.

## 9.5 Inhomogeneities to anisotropies

We now have a good handle on the perturbations to the photons at recombination. Specifically, we have predictions for $\Theta_0(\boldsymbol{k}, \eta_*)$, $\Theta_1(\boldsymbol{k}, \eta_*)$ given the initial conditions $\Phi(\boldsymbol{k}, 0)$, or equivalently $\mathcal{R}(\boldsymbol{k})$. It is time to transform this understanding into predictions for the anisotropy spectrum today. First, we will solve for the moments $\Theta_l$ at $\eta_0$ in the next subsection. Then we will spend a bit of time relating these moments to the observables. The main purpose of the following subsections is to derive Eq. (9.59), which relates the moments today to the monopole and dipole at recombination, and Eq. (9.74), which expresses the CMB power spectrum in terms of these moments today.

### 9.5.1 Free streaming

We want to derive a solution for the photon moments today $\Theta_l(k, \eta_0)$ in terms of the monopole and dipole at recombination. A formal solution can be obtained by returning to Eq. (5.67). Subtracting $\tau'\Theta$ from both sides leads to

$$\Theta' + (ik\mu - \tau')\Theta = \hat{S} \tag{9.45}$$

where the source function is defined as

$$\hat{S} \equiv -\Phi' - ik\mu\Psi - \tau'\left[\Theta_0 + \mu u_{\rm b} - \frac{1}{2}\mathcal{P}_2(\mu)\Pi\right]. \tag{9.46}$$

The left-hand side of Eq. (9.45) can be written as

$$\Theta' + (ik\mu - \tau')\Theta = e^{-ik\mu\eta+\tau}\frac{d}{d\eta}\left[\Theta e^{ik\mu\eta-\tau}\right]. \tag{9.47}$$

Using this, and multiplying both sides of Eq. (9.45) by $e^{ik\mu\eta-\tau}$ and then integrating over $\eta$ leads directly to

$$\Theta(\eta_0) = \Theta(\eta_{\rm init})e^{ik\mu(\eta_{\rm init}-\eta_0)}e^{-\tau(\eta_{\rm init})} + \int_{\eta_{\rm init}}^{\eta_0} d\eta\,\hat{S}(\eta)e^{ik\mu(\eta-\eta_0)-\tau(\eta)} \tag{9.48}$$

where we have used the fact that $\tau(\eta_0) = 0$ since $\tau$ is defined as the scattering optical depth integrated backward from today, $\eta_0$. We also know that, if the initial time $\eta_{\rm init}$ is early enough, then the optical depth $\tau(\eta_{\rm init})$ will be extremely large. Therefore, the first term on the right side of Eq. (9.48) vanishes. This corresponds to the fact that any initial anisotropy is completely erased by Compton scattering. By the same reasoning, we can set the lower limit on the integral to zero: any contribution to the integrand from $\eta < \eta_{\rm init}$ is completely

negligible. Thus, the solution for the anisotropies is

$$\Theta(k,\mu,\eta_0) = \int_0^{\eta_0} d\eta \hat{S}(k,\mu,\eta) e^{ik\mu(\eta-\eta_0)-\tau(\eta)}. \tag{9.49}$$

All the dependence on the photon direction is now in the argument $\mu$ on the right-hand side. So all of the complication is hidden in the source function $\hat{S}$. Let us forget for now the dependence of $\hat{S}$ on the angle $\mu$. We can immediately turn Eq. (9.49) into an equation for each of the $\Theta_l$: multiply each side by the Legendre polynomial $\mathcal{P}_l(\mu)$ and then integrate over all $\mu$. Via Eq. (5.66), the left-hand side gives $(-i)^l\Theta_l$, while on the right we have the integral

$$\int_{-1}^{1} \frac{d\mu}{2} \mathcal{P}_l(\mu) e^{ik\mu(\eta-\eta_0)} = \frac{1}{(-i)^l} j_l\left[k(\eta-\eta_0)\right] \tag{9.50}$$

where $j_l$ is the spherical Bessel function. This approach will in fact lead us to success, not least because the $\mu$-dependence of $\hat{S}$ is quite simple (in the end, this is due to tight coupling before recombination). So let us proceed to obtain the expression for $\Theta_l$,

$$\Theta_l(k,\eta_0) = (-1)^l \int_0^{\eta_0} d\eta \hat{S}(k,\eta) e^{-\tau(\eta)} j_l[k(\eta-\eta_0)] \qquad \text{(no } \mu\text{-dependence)}. \tag{9.51}$$

Now we can account for the $\mu$ dependence in $\hat{S}$ by noting that $\hat{S}$ multiplies the exponential $e^{ik\mu(\eta-\eta_0)}$ in Eq. (9.49). Thus, everywhere we encounter a factor of $\mu$ in $\hat{S}$ we can replace it with a time derivative:

$$\mu \to \frac{1}{ik}\frac{d}{d\eta}. \tag{9.52}$$

Then, through integration by parts, we turn each power of $\mu$ into a derivative operator acting on the remainder of the integrand, which then allows us to perform the Legendre decomposition. Let us demonstrate this explicitly with the $-ik\mu\Psi$ term in $\hat{S}$. The integral is

$$\begin{aligned} -ik\int_0^{\eta_0} d\eta\, \mu\Psi e^{ik\mu(\eta-\eta_0)-\tau(\eta)} &= -\int_0^{\eta_0} d\eta \Psi e^{-\tau(\eta)} \frac{d}{d\eta} e^{ik\mu(\eta-\eta_0)} \\ &= \int_0^{\eta_0} d\eta e^{ik\mu(\eta-\eta_0)} \frac{d}{d\eta}\left[\Psi e^{-\tau(\eta)}\right] \end{aligned} \tag{9.53}$$

where the last line follows by integration by parts. Note that the surface terms can be dropped: at $\eta = 0$ they are set to zero by the $e^{-\tau(0)}$ factor. The terms at $\eta = \eta_0$ are not small, but they are irrelevant since they have no $\mu$ dependence. They alter the observed monopole of the CMB, an alteration which we cannot detect. In other words, we can absorb them in the definition of the mean CMB temperature $T_0$. Thus, accounting for the integration by parts changes the substitution rule of Eq. (9.52) by a minus sign, with the

understanding that the derivative does *not* act on the oscillating part of the exponential, $e^{ik\mu(\eta-\eta_0)}$. The solution in Eq. (9.51) therefore becomes

$$\Theta_l(k,\eta_0) = \int_0^{\eta_0} d\eta S(k,\eta)\, j_l\left[k(\eta_0-\eta)\right] \tag{9.54}$$

with the source function now defined as

$$\begin{aligned} S(k,\eta) \equiv e^{-\tau}&\left[-\Phi' - \tau'\left(\Theta_0 + \frac{1}{4}\Pi\right)\right] \\ &+ \frac{d}{d\eta}\left[e^{-\tau}\left(\Psi - \frac{iu_{\rm b}\tau'}{k}\right)\right] - \frac{3}{4k^2}\frac{d^2}{d\eta^2}\left[e^{-\tau}\tau'\Pi\right]. \end{aligned} \tag{9.55}$$

In Eq. (9.54), we have also used the even/odd property of spherical Bessel functions: $j_l(x) = (-1)^l j_l(-x)$.

At this stage, it is useful to introduce the *visibility function*

$$g(\eta) \equiv -\tau'(\eta)e^{-\tau(\eta)}. \tag{9.56}$$

The integral $\int_0^{\eta_0} d\eta g(\eta) = 1$, so we can think of it as a probability density: $g(\eta)$ is the probability that a photon last scattered at $\eta$. Since $\tau$ is so large early on, this probability is essentially zero for $\eta$ earlier than the time of recombination. It also declines rapidly after recombination, because the prefactor $-\tau'$, which is the scattering rate, is quite small. Fig. 9.9 shows the visibility function in the fiducial cosmology.

The source function in Eq. (9.55) can now be expressed in terms of the visibility function. If we drop the polarization tensor $\Pi$ in the source since it is very small, then the source function becomes

$$\begin{aligned} S(k,\eta) \simeq\, & g(\eta)\left[\Theta_0(k,\eta) + \Psi(k,\eta)\right] \\ &+ \frac{i}{k}\frac{d}{d\eta}\left[u_{\rm b}(k,\eta)g(\eta)\right] \\ &+ e^{-\tau}\left[\Psi'(k,\eta) - \Phi'(k,\eta)\right]. \end{aligned} \tag{9.57}$$

We can take our analytic solution one step further by performing the time integral in Eq. (9.54). The source term proportional to $u_{\rm b}$ is best treated by integrating by parts. Then,

$$\begin{aligned} \Theta_l(k,\eta_0) = &\int_0^{\eta_0} d\eta\; g(\eta)\left[\Theta_0(k,\eta) + \Psi(k,\eta)\right] j_l\left[k(\eta_0-\eta)\right] \\ &- \frac{i}{k}\int_0^{\eta_0} d\eta\; g(\eta)u_{\rm b}(k,\eta)\frac{d}{d\eta}j_l\left[k(\eta_0-\eta)\right] \\ &+ \int_0^{\eta_0} d\eta\; e^{-\tau}\left[\Psi'(k,\eta) - \Phi'(k,\eta)\right]\; j_l\left[k(\eta_0-\eta)\right]. \end{aligned} \tag{9.58}$$

There are two types of terms in Eq. (9.58). The integrals in the first two lines are weighted by the visibility function. These are the dominant terms. The integral in the last line, on the

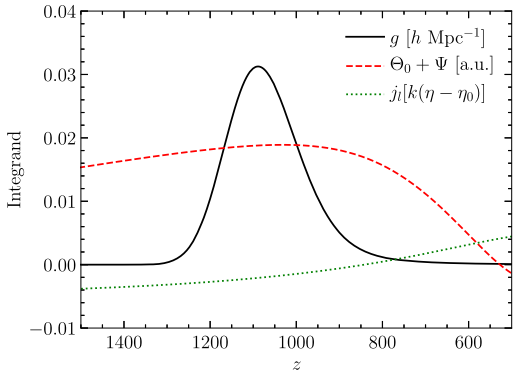



**FIGURE 9.9** The visibility function $g(\eta)$ (black solid) in comparison with the other two components of the integrand in the first line of Eq. (9.58). The visibility function is sharply peaked, so it changes rapidly compared with the monopole $\Theta_0 + \Psi$ (shown in arbitrary units) and the Bessel function $j_l(k[\eta - \eta_0])$. Results shown are for $l = 220$, $k = 0.02$ Mpc$^{-1}$, corresponding to the first peak in the CMB anisotropy spectrum.

other hand, is weighted by $e^{-\tau}$, so the integrand contributes as long as $\tau \lesssim 1$, that is, at all times after recombination. Note that, if the potentials are constant after recombination, which is the case during matter domination, the terms in the last line vanish.

Since the visibility function is so sharply peaked, the integrals in the first two terms become very simple. To see why, consider Fig. 9.9 which shows the three parts of the integrand of the first term (the "monopole") in Eq. (9.58). Since the visibility function changes rapidly compared with the other two functions, we can evaluate those other functions at the peak of the visibility function, i.e., at $\eta = \eta_*$, and remove them from the integral. But then, the integral is simply $\int d\eta g(\eta) = 1$. Thus, we are left with

$$\begin{aligned}\Theta_l(k,\eta_0) \simeq & \left[\Theta_0(k,\eta_*) + \Psi(k,\eta_*)\right]\, j_l\left[k(\eta_0-\eta_*)\right] \\ & + 3\Theta_1(k,\eta_*)\left(j_{l-1}\left[k(\eta_0-\eta_*)\right] - (l+1)\frac{j_l\left[k(\eta_0-\eta_*)\right]}{k(\eta_0-\eta_*)}\right) \\ & + \int_0^{\eta_0} d\eta\, e^{-\tau}\left[\Psi'(k,\eta) - \Phi'(k,\eta)\right]\, j_l\left[k(\eta_0-\eta)\right]. \end{aligned} \tag{9.59}$$

Here we have used the spherical Bessel function identity of Eq. (C.19) to rewrite the Bessel function derivative in the velocity term and also the fact that $u_{\rm b} \simeq -3i\Theta_1$ at $\eta_*$. On scales much smaller than the one shown in Fig. 9.9, $\Theta_0 + \Psi$ changes more rapidly because of the rapid change in the damping scale around recombination (see Fig. 9.8). Incorporating damping by multiplying $(\Theta_0 + \Phi)(k,\eta_*)$ by $e^{-k^2/k_D^2(\eta_*)}$ does not capture this accurately. A much better approximation of the rapid changes in the damping scale is to change the multiplication factor

$$e^{-k^2/k_D^2(\eta_*)} \to \int d\eta g(\eta) e^{-k^2/k_D^2(\eta)}. \tag{9.60}$$

Eq. (9.59) is the basis for semi-analytic calculations (Seljak, 1994; Hu and Sugiyama, 1995) of anisotropy spectra that agree with the numerical solutions to within $\sim 10\%$. From Eq. (9.59), we see that, to solve for the anisotropies today, we must know the monopole ($\Theta_0$), dipole ($\Theta_1$), and potential ($\Psi$) at the time of recombination. Further, there will be corrections if the potentials are time dependent. These corrections, encoded in the last line of Eq. (9.59), are called *integrated Sachs–Wolfe* terms.

The monopole term in Eq. (9.59) is precisely what we expected from the rough arguments of Sect. 9.1. First, it involves the combination $\Theta_0 + \Psi$ of the temperature anisotropy and gravitational redshift. Second, the spherical Bessel function, $j_l[k(\eta_0 - \eta_*)]$, determines how much anisotropy on an angular scale $l^{-1}$ is contributed by a plane wave with wavenumber $k$. On very small angular scales.

$$j_l(x) \stackrel{x/l\to 0}{\longrightarrow} \frac{1}{l}\left(\frac{x}{l}\right)^{l-1/2}. \tag{9.61}$$

That is, $j_l(x)$ is extremely small for large $l$ when $x < l$. In our case, this means that $\Theta_l(k, \eta_0)$ is very close to zero for $l > k\eta_0$. This makes sense physically. Returning to Fig. 9.5, we see that very small angular scales will see little anisotropy from a perturbation with a large wavelength. The converse is also true: angular scales larger than $1/(k\eta_0)$ get little contribution from such a perturbation (see also Fig. 9.13). To sum up, a perturbation with wavenumber $k$ contributes predominantly on angular scales of order $l \sim k\eta_0$.

### 9.5.2 The angular power spectrum

How is the observed anisotropy pattern today related to the rather abstract $\Theta_l(k, \eta_0)$, which refer to a plane-wave perturbation $\boldsymbol{k}$? To answer this question, we must first describe the way in which the temperature field is characterized today and then relate this characterization to $\Theta_l$.

Recall that in Eq. (5.2), we wrote the temperature of the CMB radiation field in the universe as

$$T(\boldsymbol{x}, \hat{\boldsymbol{p}}, \eta) = T(\eta)\left[1 + \Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, \eta)\right]. \tag{9.62}$$

Although this field is defined at every point in space and time, we can observe it only here (at $\boldsymbol{x}_0$) and now (at $\eta_0$).[3] Our only handle on the anisotropies is their dependence on the direction of the incoming photons, $\hat{\boldsymbol{p}}$. So all the richness we observe comes from the changes in the temperature as the direction vector $\hat{\boldsymbol{p}}$ changes. Observers typically make maps, wherein the temperature is reported at a number of incoming directions, or "locations on the sky." These locations are usually labeled not by the $\hat{p}_x$, $\hat{p}_y$, $\hat{p}_z$ components of $\hat{\boldsymbol{p}}$, but rather by polar coordinates $\theta$, $\phi$. However, it is a simple matter to move back and

[3] We do make small excursions from this point in spacetime. For example, satellites are not located on Earth and anisotropy measurements have been made over the past 30 years. These excursions are completely insignificant on scales over which the temperature is varying, however, which are of order tens of Mpc and the Hubble time, respectively.

forth between the unit vector $\hat{\boldsymbol{p}}$ and polar coordinates. We will continue to use the by-now familiar $\hat{\boldsymbol{p}}$ in the derivation.

We now expand the temperature perturbation in terms of spherical harmonics. That is, we write

$$\Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, \eta) = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} a_{lm}(\boldsymbol{x}, \eta) Y_{lm}(\hat{\boldsymbol{p}}). \tag{9.63}$$

The subscripts $l$, $m$ are conjugate to the real-space unit vector $\hat{\boldsymbol{p}}$, just as the variable $\boldsymbol{k}$ is conjugate to the three-dimensional position $\boldsymbol{x}$. We are by now familiar with Fourier transforms, so it is useful to think of the expansion in terms of spherical harmonics as a kind of 2D Fourier transform. Whereas the complete set of eigenfunctions for the 3D Fourier transform are $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, here the complete set of eigenfunctions for expansion on the surface of a sphere are $Y_{lm}(\hat{\boldsymbol{p}})$ (see Appendix C.2). All of the information contained in the temperature field $T$ is also contained in the $(\boldsymbol{x}, \eta)$-dependent amplitudes $a_{lm}$. As an example of this, consider an experiment that maps the full sky with an angular resolution of 7°. The full sky has $4\pi$ radians$^2 \simeq 41{,}000$ degrees$^2$, so there are 840 pixels with area of $(7°)^2$. Thus, such an experiment would have 840 independent pieces of information. Were we to characterize this information with $a_{lm}$ instead of temperatures in pixels, there would be some $l_{\rm max}$ above which there is no information. One way to determine this $l_{\rm max}$ is to set the total number of recoverable $a_{lm}$ as $\sum_{l=0}^{l_{\rm max}}(2l+1) = (l_{\rm max}+1)^2 = 840$. So the information could be equally well characterized by specifying all the $a_{lm}$ up to $l_{\rm max} = 28$. Incidentally, this is a fairly good caricature of the COBE experiment (Smoot et al., 1992; Bennett et al., 1996), the discoverer of CMB anisotropies. They presented temperature data over many more pixels, but these pixels were overlapping. So, the independent information was contained in multipoles up to $l \sim 30$. The current generation of experiments is capable of measuring the moments all the way up to $l$ of several thousands, at which point the primary CMB anisotropies are sufficiently damped by photon diffusion that the observed radiation is dominated by astrophysical foreground sources (as well as gravitational lensing, Sect. 13.3).

We want to relate the observables, the $a_{lm}$, to the moments of the temperature distribution we have been dealing with. To do this, we can use the orthogonality property of the spherical harmonics. The $Y_{lm}$ are normalized via Eq. (C.11),

$$\int d\Omega\, Y_{lm}(\hat{\boldsymbol{p}}) Y^*_{l'm'}(\hat{\boldsymbol{p}}) = \delta_{ll'}\delta_{mm'}. \tag{9.64}$$

Therefore the expansion of $\Theta$ in terms of spherical harmonics, Eq. (9.63), can be inverted by multiplying both sides by $\mathrm{Y}^*_{lm}(\hat{\boldsymbol{p}})$ and integrating:

$$a_{lm}(\boldsymbol{x}, \eta) = \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \int d\Omega\, Y^*_{lm}(\hat{\boldsymbol{p}}) \Theta(\boldsymbol{k}, \hat{\boldsymbol{p}}, \eta). \tag{9.65}$$

Here we have written the right-hand side in terms of the Fourier transform ($\Theta(\boldsymbol{k})$ instead of $\Theta(\boldsymbol{x})$), since that is the quantity for which we obtained solutions.

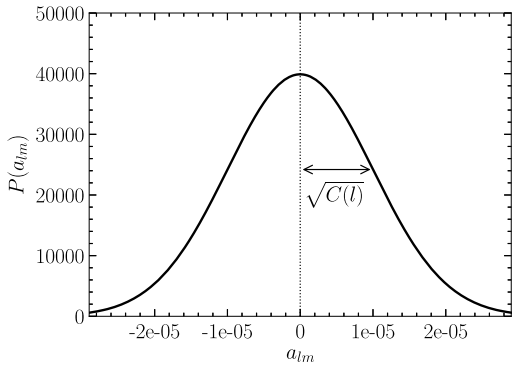



**FIGURE 9.10** The distribution from which the $a_{lm}$ are drawn. The Gaussian distribution has expectation value equal to zero and an RMS width of $[C(l)]^{1/2}$.

As with the density perturbations, we cannot make predictions about any particular $a_{lm}$, just about the distribution from which they are drawn, a Gaussian distribution that traces its origin to the quantum fluctuations laid down during inflation. Fig. 9.10 illustrates this Gaussian distribution. The mean value of the $a_{lm}$ is zero, but they have a nonzero variance. The variance of the $a_{lm}$ is called $C(l)$. Thus,

$$\langle a_{lm} \rangle = 0; \qquad \langle a_{lm} a^*_{l'm'} \rangle = \delta_{ll'} \delta_{mm'} C(l), \tag{9.66}$$

where $\langle \cdot \rangle$ now denotes an ensemble average, i.e. the result obtained in the limit of measuring over an infinite volume. It is very important to note that, for a given $l$, each $a_{lm}$ has the same variance. For $l = 100$, say, all 201 $a_{100,m}$ are drawn from the same distribution. When we measure these 201 coefficients, we are sampling the distribution. These 201 samples give us a good handle on the underlying variance of the distribution (this will be made more rigorous in Sect. 14.1). On the other hand, if we measure the five components of the quadrupole ($l = 2$), we have much less statistical precision on the underlying variance, $C(2)$. Thus, *there is a fundamental uncertainty in the knowledge we may get about the* $C(l)$. This uncertainty, which is most pronounced at low $l$, is called *cosmic variance*. Quantitatively, the uncertainty scales as the inverse of the square root of the number of samples. More precisely, it is the uncertainty on the estimate of $C(l)$ after using the $2l + 1$ samples to infer it (again, more on this in Ch. 14):

$$\left( \frac{\Delta C(l)}{C(l)} \right)_{\text{cosmic variance}} = \sqrt{\frac{2}{2l+1}}. \tag{9.67}$$

In practice, this limit is never quite achieved, because even if an instrument observes the full sky (such as the satellite experiments COBE, WMAP, and Planck), the large foreground emission in the Milky Way plane means that some parts of the sky need to be masked. For a measurement based on a fraction $f_{\text{sky}}$ of the full sky, the error bar is increased by roughly a factor $1/\sqrt{f_{\text{sky}}}$.

We can now obtain an expression for $C(l)$ in terms of $\Theta_l(k)$. You might rightfully worry about a notation collision here, but we will see that the index $l$ in $C(l)$ and $\Theta_l$ is indeed the same. First we square $a_{lm}$ in Eq. (9.65) and take the expectation value of the distribution. For this we need $\langle\Theta(\boldsymbol{k},\hat{\boldsymbol{p}})\Theta^*(\boldsymbol{k}',\hat{\boldsymbol{p}}')\rangle$, where from now on we will keep the $\eta=\eta_0$ dependence implicit. This expectation value is complicated because it depends on two separate phenomena: $(i)$ the initial amplitude and phase of the perturbation, randomly chosen during inflation from a Gaussian distribution, and $(ii)$ the evolution we have studied in this chapter that turns this initial perturbation into anisotropies. The former is a random variable, the latter is deterministic: given an initial value for the amplitude and phase of the perturbation, the equations uniquely determine its evolution. To simplify then, it makes sense to separate these two phenomena and write the photon distribution as $\mathcal{R}\times(\Theta/\mathcal{R})=\mathcal{R}\times\mathcal{T}$, where the primordial curvature perturbation $\mathcal{R}$ depends on $\boldsymbol{k}$, but not on the direction vector $\hat{\boldsymbol{p}}$. The ratio

$$\mathcal{T}(\boldsymbol{k},\hat{\boldsymbol{p}})\equiv\frac{\Theta(\boldsymbol{k},\hat{\boldsymbol{p}},\eta_0)}{\mathcal{R}(\boldsymbol{k})} \tag{9.68}$$

is precisely what we have solved for in this chapter: given the initial amplitude of a mode, we have learned how to evolve forward in time. $\mathcal{T}(\boldsymbol{k},\hat{\boldsymbol{p}})$ does *not* depend on the initial amplitude of each mode and is not random, so it can be removed from the averaging over the distribution. Therefore,

$$\begin{aligned}\langle\Theta(\boldsymbol{k},\hat{\boldsymbol{p}})\Theta(\boldsymbol{k}',\hat{\boldsymbol{p}}')\rangle&=\langle\mathcal{R}(\boldsymbol{k})\mathcal{R}^*(\boldsymbol{k}')\rangle\,\mathcal{T}(\boldsymbol{k},\hat{\boldsymbol{p}})\mathcal{T}^*(\boldsymbol{k}',\hat{\boldsymbol{p}}')\\&=(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}')\,P_{\mathcal{R}}(k)\,\mathcal{T}(\boldsymbol{k},\hat{\boldsymbol{p}})\mathcal{T}^*(\boldsymbol{k},\hat{\boldsymbol{p}}'),\end{aligned} \tag{9.69}$$

where the second equality uses the definition of the power spectrum of curvature perturbations $P_{\mathcal{R}}(k)$. Now we make use of one more simplification, which holds specifically for the scalar perturbations we have focused on: the ratio (or transfer function) $\mathcal{T}$ only depends on $\hat{\boldsymbol{p}}$ through its angle with $\hat{\boldsymbol{k}}$, $\mu=\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}}$:

$$\mathcal{T}(\boldsymbol{k},\hat{\boldsymbol{p}})=\mathcal{T}(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}}). \tag{9.70}$$

This will help us do the angular integrals in the following. After squaring Eq. (9.65), the anisotropy spectrum becomes

$$C(l)=\int\frac{d^3k}{(2\pi)^3}P_{\mathcal{R}}(k)\int d\Omega Y^*_{lm}(\hat{\boldsymbol{p}})\mathcal{T}(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}})\int d\Omega' Y_{lm}(\hat{\boldsymbol{p}}')\mathcal{T}^*(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}}'). \tag{9.71}$$

Now we can expand $\mathcal{T}(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}})$ and $\mathcal{T}(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}}')$ in Legendre polynomials using the inverse of Eq. (5.66):

$$\mathcal{T}(k,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}})=\sum_l(-i)^l(2l+1)\mathcal{P}_l(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{p}})\mathcal{T}_l(k). \tag{9.72}$$

So, $\mathcal{T}_l(k) = \Theta_l(k, \eta_0)/\mathcal{R}(\boldsymbol{k})$. This leaves

$$C(l) = \int \frac{d^3k}{(2\pi)^3} P_{\mathcal{R}}(k) \sum_{l'l''} (-i)^{l'} (i)^{l''} (2l'+1)(2l''+1) \mathcal{T}_{l'}(k) \mathcal{T}^*_{l''}(k) \\ \times \int d\Omega \mathcal{P}_{l'}(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{p}}) Y^*_{lm}(\hat{\boldsymbol{p}}) \int d\Omega' \mathcal{P}_{l''}(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{p}}') Y_{lm}(\hat{\boldsymbol{p}}'). \tag{9.73}$$

The two angular integrals here (Exercise 9.9) are identical. They are nonzero only if $l' = l$ and $l'' = l$, in which case they are equal to $4\pi Y_{lm}(\hat{\boldsymbol{k}})/(2l+1)$ and its complex conjugate, respectively. The angular part $d\Omega$ of the integral over $\boldsymbol{k}$ then becomes an integral over $|Y_{lm}|^2$, which is just equal to 1, leaving

$$C(l) = \frac{2}{\pi} \int_0^\infty dk\, k^2\, P_{\mathcal{R}}(k)\, |\mathcal{T}_l(k)|^2 . \tag{9.74}$$

For a given $l$, then, the variance $C(l)$ is an integral over all Fourier modes of the variance of $\Theta_l(\boldsymbol{k})$, which is given by $|\mathcal{T}_l(k)|^2$ times the variance of curvature perturbations. We can then use Eq. (9.59) and Eq. (9.74) to compute the anisotropy spectrum today.

As an example, we can rewrite the tight-coupling solution of Eq. (9.26) as

$$\Theta_0(\boldsymbol{k}, \eta) = \\ \mathcal{R}(\boldsymbol{k}) \left[ -\frac{2}{3} \frac{\Phi(\boldsymbol{k}, \eta)}{\Phi(\boldsymbol{k}, 0)} + \cos(k r_s) + \frac{4}{3} \frac{k}{\sqrt{3}} \int_0^\eta d\tilde{\eta} \frac{\Phi(\boldsymbol{k}, \tilde{\eta})}{\Phi(\boldsymbol{k}, 0)} \sin[k(r_s(\eta) - r_s(\tilde{\eta}))] \right] e^{-k^2/k_D^2(\eta)}. \tag{9.75}$$

The exponential factor at the end accounts for damping. This works similarly for the dipole in Eq. (9.28). Then, we insert these expressions into Eq. (9.59), which evolves the anisotropies forward using free-streaming, to obtain $\mathcal{T}_l(k) = \Theta_l(k, \eta_0)/\mathcal{R}(\boldsymbol{k})$.

## 9.6 The CMB power spectrum

### 9.6.1 Large angular scales

The large-angle CMB anisotropies are determined by extremely large-scale modes that have entered our horizon only recently. As such, they offer a particularly direct way of measuring the initial conditions. On these largest of scales, we can neglect the dipole in Eq. (9.59). So the large-angle anisotropy is determined by $\Theta_0 + \Psi$ evaluated at recombination, in addition to the last term in Eq. (9.59) which we will turn to below. The large-scale solution we found in Eq. (9.6) was that the combination $\Theta_0 + \Psi$ is equal to $-\mathcal{R}/5$. This gives us what we need: an expression for $\Theta_0 + \Psi$ at recombination that we can plug into the monopole term in Eq. (9.59). To get the anisotropy spectrum today, we then integrate as in Eq. (9.74), leaving

$$C(l)^{\mathrm{SW}} \simeq \frac{2}{25\pi} \int_0^\infty dk\, k^2\, P_{\mathcal{R}}(k) \left| j_l \left[ k(\eta_0 - \eta_*) \right] \right|^2 \tag{9.76}$$

where the superscript denotes *Sachs–Wolfe*, in honor of the authors of the first paper to compute the large-angle anisotropy (Sachs and Wolfe, 1967). The power spectrum of curvature perturbations is given by Eq. (7.99). Therefore,

$$C(l)^{\mathrm{SW}} \simeq \frac{4\pi}{25}\mathcal{A}_s k_{\mathrm{p}}^{1-n_s}\int_0^\infty dk\, k^{n_s-2}\, j_l^2\left[k(\eta_0-\eta_*)\right]. \tag{9.77}$$

The integral here can be computed analytically. First, we will use the fact that $\eta_* \ll \eta_0$ and define the integration variable $x \equiv k\eta_0$. Then, Eq. (9.77) can be rewritten as

$$C(l)^{\mathrm{SW}} \simeq \frac{4\pi}{25}\mathcal{A}_s(\eta_0 k_{\mathrm{p}})^{1-n_s}\int_0^\infty dx\, x^{n_s-2}\, j_l^2(x). \tag{9.78}$$

The integral over the spherical Bessel functions can be analytically expressed in terms of gamma functions (Eq. (C.18)), leaving

$$C(l)^{\mathrm{SW}} \simeq 2^{n_s-2}\frac{\pi^2}{25}\mathcal{A}_s(\eta_0 k_{\mathrm{p}})^{1-n_s}\frac{\Gamma\left(l+\frac{n_s}{2}-\frac{1}{2}\right)}{\Gamma\left(l+\frac{5}{2}-\frac{n_s}{2}\right)}\frac{\Gamma(3-n_s)}{\Gamma^2\left(2-\frac{n_s}{2}\right)}. \tag{9.79}$$

If the spectrum is scale-invariant, $n_s = 1$, then the first ratio of the gamma functions $\Gamma(l)/\Gamma(l+2)$ is equal to $[l(l+1)]^{-1}$ using Eq. (C.27). The remaining ratio of gamma functions $\Gamma(2)/\Gamma^2(3/2) = 4/\pi$ using Eq. (C.28), so

$$l(l+1)C(l)^{\mathrm{SW}} = \frac{8}{25}\mathcal{A}_s \tag{9.80}$$

is a constant. Indeed, $l(l+1)C(l)$ is the variance of the temperature anisotropies per logarithmic interval in $l$, analogously to $k^3 P_{\mathcal{R}}(k)$ for the three-dimensional power spectrum. Since the latter is a constant if $n_s = 1$, it is perhaps not surprising that $l(l+1)C(l)$ becomes a constant in this case. It has become customary to plot $l(l+1)C(l)$ vs $l$ on a logarithmic scale, which then becomes approximately constant at low $l$.

Fig. 9.11 shows the Planck measurements of the large-angular-scale anisotropies along with the Boltzmann solutions of the fiducial Euclidean ΛCDM model. The deviation from a constant is due to the ISW effect and the contribution from the dipole (neglected in Eq. (9.79)) becoming nonnegligible at higher $l$. Nevertheless, Eq. (9.80) provides a reasonable approximation. Since the $y$-axis gives the variance contributed by a given scale $l$, we can read off the amplitude of the large-angular-scale fluctuations: roughly, $\langle(\Delta T/T_0)^2\rangle \sim 10^{-10}$, so the RMS fluctuations are the square root of this, of order $10^{-5}T_0 = 27\ \mu\mathrm{K}$.

Going beyond the scale-invariant case, following our analytic result, the power spectrum multiplied by $l(l+1)$ should scale as $(l/l_{\mathrm{p}})^{n_s-1}$, where $l_{\mathrm{p}}$ is the angular wavenumber roughly corresponding to the pivot scale $k_{\mathrm{p}}$. You can see this scaling from Eq. (9.79) or more directly from the integral in Eq. (9.78). The integrand peaks at $x \sim l$, so roughly every appearance of $x$ there can be replaced by $l$. The generalization of the integrand from $x^{-1}$ to $x^{n_s-2}$ therefore leads to a change in the spectrum that scales as $l^{n_s-1}$. Given the smallness

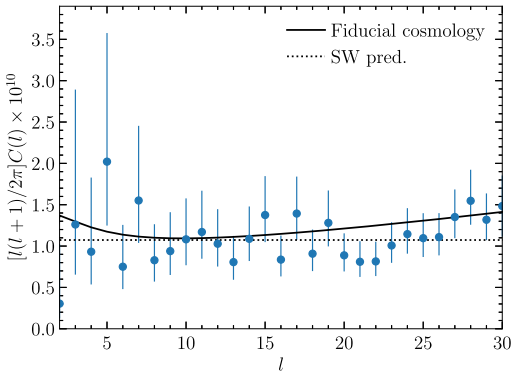



**FIGURE 9.11** Large-scale CMB power spectrum as measured by Planck (Planck Collaboration, 2018b), and fiducial ΛCDM prediction (solid). The dotted line shows the scale-invariant Sachs–Wolfe plateau predicted by Eq. (9.80).

of $n_s - 1$, this scaling is, however, masked by the other contributions mentioned above. To get constraints on the spectral index as well as the amplitude, the data have to span a larger range in $l$. That is, we have to include anisotropies on smaller scales.

## 9.6.2 Acoustic peaks

On smaller scales, i.e. those that are inside the horizon at recombination, the anisotropy spectrum depends on all terms in Eq. (9.59): the monopole $\Theta_0$, the dipole $\Theta_1$, and the integrated Sachs–Wolfe effect, $\propto \int d\eta(\Psi - \Phi)'$. Fig. 9.12 shows all these contributions to the angular power spectrum. Let us consider each in turn.

The monopole at recombination $(\Theta_0 + \Psi)(k, \eta_*)$ free-streams to us today, creating anisotropies on angular scales $l \sim k\eta_0$. This is what we expected back in Fig. 9.5, showed to be true in Eq. (9.59), and what we can now see directly in Fig. 9.12. There are two interesting features of the quantitative aspect of the free-streaming process. First, note that the "zeros" in the monopole spectrum, here at $l \sim 70$, 400, 650, and 1000, are smoothed out because many Fourier modes contribute to anisotropy on a given angular scale. If only the $k = 400/\eta_0$ modes contributed to the anisotropy at $l = 400$, then $C(400)$ would really be zero. But many nonzero modes, with wavenumbers different from $400/\eta_0$, contribute. These change the zero to a trough in the $C(l)$ spectrum.

The second feature of free-streaming worth noticing is that our initial estimate of the peak positions is not exactly right. Inhomogeneity on scale $k$ does *not* show up as anisotropy precisely on angular scale $l = k\eta_0$. Rather, there is a noticeable shift, suggesting that a given $k$-mode contributes to slightly smaller $l$ than we anticipated. This shift partially arises from the spherical Bessel function in Eq. (9.59). As shown in Fig. 9.13, the peak in the Bessel function comes not when $l = k\eta_0$, but rather at slightly smaller values of $l$. A better approximation for the first peak position is $l_{\rm pk} \simeq 0.75\pi\eta_0/r_s$.

The dipole at recombination is smaller than the monopole and out of phase with it. The dashed line in Fig. 9.12 shows that the effect of adding it is to raise the overall anisotropy

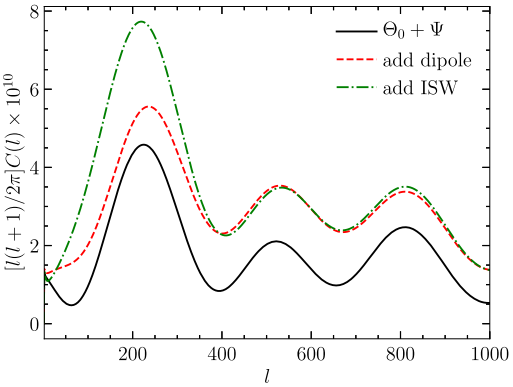



**FIGURE 9.12** Intermediate- to small-scale CMB power spectrum. The solid black line shows the result obtained if only the monopole at recombination $(\Theta_0 + \Psi)(k = l/\eta_0, \eta_*)$ were present, and contains most of the structure of the final anisotropy spectrum. Including the dipole (red dashed) raises the anisotropy spectrum. Since the dipole is out of phase with the monopole, the troughs become less pronounced. Adding the integrated Sachs–Wolfe effect (green dash-dotted) enhances the anisotropy mostly on scales comparable to or larger than the horizon at recombination. Thus the first peak gets most of the additional power.

level, but in particular in the region of the troughs, lowering the prominence of the peaks. This is a direct manifestation of the dipole and monopole being out of phase with one another. That is, at the places where the monopole contributes least to the anisotropies, at its troughs, the dipole contributes the most. Another feature of the monopole and dipole contributions is that they add incoherently. By incoherently, we mean that the cross-term of $\Theta_l$ from the monopole multiplied by $\Theta_l$ from the dipole vanishes when integrating over all $k$-modes to get the $C(l)$. This can be seen mathematically from the properties of the spherical Bessel function (Exercise 9.11). Incoherence implies that the dipole is not as important in the power spectrum as one might naively think. If the amplitude of the dipole is 30% of that of the monopole at recombination, the dipole's contribution to the $C(l)$ is only 10%.

The third contribution is from the integrated Sachs–Wolfe effect due to the time evolution of the potentials after recombination, which is mostly due to the fact that the energy density in radiation is not entirely negligible at recombination. If the universe were purely matter dominated, there would be no such effect. But, the transition to pure matter domination is not abrupt, and even for $a_{\rm eq} \sim 10^{-4}$, an ISW effect occurs right after recombination. To see which scales are affected by the ISW effect, consider the integral in Eq. (9.59). Suppose the potential evolves at time $\eta_c$, with all sub-horizon scales $(k\eta_c > 1)$ being affected. The Bessel function peaks at $l \sim k(\eta_0 - \eta_c)$; so all angular scales $l > (\eta_0 - \eta_c)/\eta_c$ are affected. The largest effect is typically on scales of the horizon at the time $\eta_c$.

This early ISW effect is particularly important because it adds coherently with the monopole. To see this, integrate the last term in Eq. (9.59) by parts. Then, the dominant contribution comes from $\eta \simeq \eta_*$, so the Bessel function can be evaluated there, leaving the

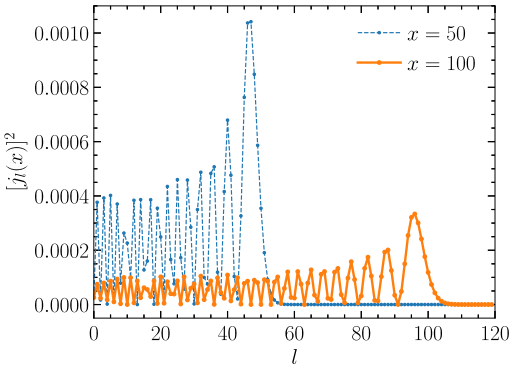



**FIGURE 9.13** The spherical Bessel function squared at $x = 50$ and $x = 100$ as a function of $l$. Note that the peak occurs when $l$ is slightly smaller than $x$.

trivial integral which gives

$$\Theta_l(k,\eta_0)^{\text{early ISW}} = \left[\Psi(k,\eta_0) - \Psi(k,\eta_*) - \Phi(k,\eta_0) + \Phi(k,\eta_*)\right] j_l\left[k(\eta_0 - \eta_*)\right]. \tag{9.81}$$

This adds exactly in phase with the monopole (which is proportional to the same Bessel function) so even though the magnitude of the effect on $\Theta_l$ is much smaller than is the dipole, the effect on the anisotropy spectrum is comparable. A 30% dipole leads to a 10% shift in the $C(l)$, while a 5% ISW effect leads to the same 10% shift in the $C(l)$. The dash-dotted line in Fig. 9.12 shows that the anisotropies on large scales, those with $l \lesssim \eta_0/\eta_*$, get a big boost from this early ISW effect.

The late-time ISW effect occurs when potentials decay during the dark energy epoch at $z \lesssim 1$ (Sect. 8.5). This late-time effect therefore is restricted to extremely large scales, $l \lesssim 30$, and is just barely visible when plotting $C(l)$ on a linear scale in $l$; it is much more obvious as the upturn at the lowest $l$ in Fig. 9.11. The most direct way to detect this effect is to cross-correlate the large-angle CMB anisotropies with large-scale structure at low redshifts, such as angular galaxy correlations (Sect. 11.2).

## 9.7 Cosmological parameters

The power spectrum of CMB anisotropies has rich structure, and its shape depends on cosmological parameters. By measuring it precisely, we can constrain the various parameters that describe the ingredients that enter the calculation. The price of this multidimensional parameter space is that there are partial degeneracies: the effect of varying one parameter can be mimicked by varying, in general, several other parameters in specific ways. In this section, we will try to get a feel for which parameters can be constrained directly, and which important degeneracies exist and how they work.

One very important decision that must be made is which parameters will be allowed to vary. We will consider the following seven ΛCDM parameters:

- Curvature parameter, $\Omega_{\rm K} \equiv 1 - \Omega_{\rm m} - \Omega_\Lambda$, often set to zero in the concordance model
- Cosmological constant, parametrized by $\Omega_\Lambda$
- Normalization of the primordial spectrum, $\mathcal{A}_s$
- Scalar spectral index, $n_s$
- Reionization, parametrized by the optical depth $\tau_{\rm rei}$ to a redshift after recombination is completed
- Baryon density, $\Omega_{\rm b}h^2$
- CDM density, $\Omega_{\rm c}h^2$.

There are two aspects of this list worth stressing. The first is that obviously it does not include all possible cosmological parameters. Some favorites missing are neutrino masses (we will set the sum of neutrino masses to its minimum experimentally allowed value, $\sum m_\nu = 0.06$ eV), the equation of state for dark energy $w$ (fixed here at $-1$ corresponding to a cosmological constant), and tensor modes (amplitude $r$ fixed to zero). The main reason for these omissions is that these parameters are not directly constrained by the CMB temperature power spectrum. The effect of neutrino masses is simply too small at early times, and the same holds for the details of dark energy (which mostly affects the CMB through the distance to the last-scattering surface). Dark energy and neutrino masses are best constrained by combining the CMB with large-scale structure probes, which are the topic of Ch. 11. On the other hand, tensor modes are now most constrained by CMB *polarization*, so we defer their discussion to Ch. 10.

The second important point is that we have deliberately chosen specific combinations of some of these parameters, e.g., $\Omega_{\rm b}h^2$, not $\Omega_{\rm b}$ and $h$ separately. Notice that, by allowing for $\Omega_{\rm K}$ and $\Omega_\Lambda$ in addition to $\Omega_{\rm m}h^2 = (\Omega_{\rm b} + \Omega_{\rm c})h^2$, we have effectively accounted for the Hubble parameter, since $\Omega_{\rm m}$ is fixed by the constraint $\Omega_{\rm m} = 1 - \Omega_\Lambda - \Omega_{\rm K}$. There is a good reason for considering this combination of parameters. $\Omega_{\rm m}h^2$, for example, parametrizes the proper physical matter density in the universe today, written in some odd units involving $8\pi G$ and 100 km/s/Mpc. The physics determining the CMB anisotropies cares much more about the physical matter density than the density parameter $\Omega_{\rm m}$. The same holds for the baryons. Finally, the physical energy density in photons, $\Omega_\gamma h^2$, is extremely well determined through the CMB temperature. Thus, the equality epoch $a_{\rm eq}$ is essentially a function of $\Omega_{\rm m}h^2$ only.

Let us now consider the effect of each parameter in turn.

### 9.7.1 Curvature and $\Lambda$

If the universe is not Euclidean, then the simple picture of Fig. 9.5 is no longer accurate, since the photon geodesics which start out parallel to each other converge or diverge. Consider the implication of this effect for anisotropies. Suppose the identical pattern of inhomogeneities was in place at recombination in both a Euclidean and an open universe (a very good approximation if $\Omega_{\rm K}$ is small). As shown in Fig. 9.14, a fixed physical scale such as that of the first peak, say, gets projected onto a much smaller angular scale in an open

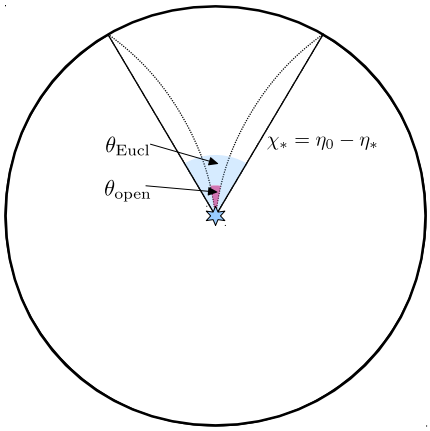



**FIGURE 9.14** In comoving coordinates $(\eta, \boldsymbol{x})$, photon trajectories in a Euclidean universe are straight lines (solid), while those in an open universe diverge (dashed). Perturbations at last scattering appear on smaller scales in an open universe ($\theta_{\rm open}$) than they do in a Euclidean universe ($\theta_{\rm Eucl}$).

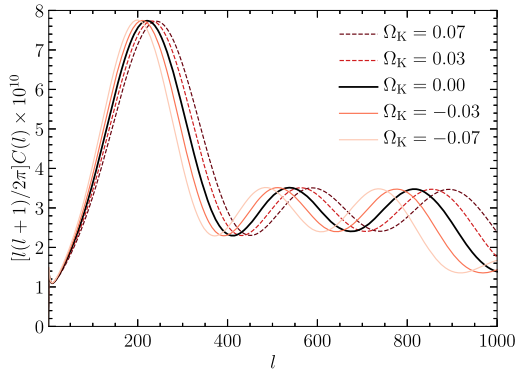



**FIGURE 9.15** The anisotropy spectrum in Euclidean versus open and closed universes. The pattern of peaks and troughs persists in curved universes but is shifted to smaller scales for a open universes ($\Omega_{\rm K} > 0$), while the opposite happens for closed universes ($\Omega_{\rm K} < 0$). Only $\Omega_{\rm K}$ and $\Omega_\Lambda$ are varied in this figure, while all other parameters are fixed at their values for the fiducial cosmology.

universe. The peaks therefore shift to higher $l$. The opposite happens in a closed universe. As shown in Fig. 9.15, this is precisely what happens in the numerical calculation.

The magnitude of this effect is determined by the angular diameter distance to the last-scattering surface, given in a Euclidean universe simply by $\eta_0 - \eta_*$, and in a universe with curvature by Eq. (2.39). Because of the large distance to last scattering, the CMB peaks respond very sensitively to curvature, resulting in a correspondingly tight constraint. Current best constraints on $\Omega_{\rm K}$, obtained by combining CMB and large-scale structure probes, are at the level of $|\Omega_{\rm K}| < 0.002$ (Planck Collaboration, 2018b). We have come a long way since the time when the open CDM model with $\Omega_{\rm K} = 1 - \Omega_{\rm m} \simeq 0.7$ was a viable scenario!

Now, an exactly Euclidean universe is only one point in parameter space, the point at which the sum of the energy densities exactly equals the critical density, and no data will ever rule out all values except for this one point. In fact, we expect to observe very small but nonzero curvature even in the inflationary paradigm. Inflation produces perturbations on all scales, including those just at our current horizon. The isotropic part of such a horizon-scale perturbation appears to us precisely as curvature, with $\Omega_K \sim (k/a_0 H_0)^2 \mathcal{R}(\boldsymbol{k})\Big|_{k=H_0}$ (this provides yet another justification for the name "curvature perturbation;" see the discussion at the end of Sect. 7.4.3). Given the approximate scale-invariant spectrum for $\mathcal{R}(\boldsymbol{k})$, inflation thus predicts that $\Omega_K$ should be a random number with RMS value of order $\sqrt{\mathcal{A}_s} \sim 10^{-4}$. Evidence for a value much larger than that would be problematic for the inflationary scenario, however.

Changing the cosmological constant has a similar effect to curvature, in that it shifts the peak locations due to the change in the angular diameter distance to last scattering (recall that we are also modifying $H_0$ when varying $\Omega_\Lambda$ while keeping $\Omega_m h^2$ fixed). After all, both are late-time phenomena that do not play a role at recombination. You will show in Exercise 9.12 that the effect of changing $\Lambda$ can be readily explained in this way. This also explains why the CMB constraint on $\Omega_K$ is partially degenerate with that on $\Omega_\Lambda$, if no large-scale structure probes are included to break the degeneracy. In addition, changing $\Lambda$ also affects the late-time ISW contribution at $l \lesssim 30$, with an increase in $\Lambda$ boosting the $C(l)$ on these scales (see Fig. 9.11), although the constraining power of this effect is unfortunately limited due to the large cosmic variance errors.

## 9.7.2 Amplitude, spectral index, and optical depth

The effect of changing the amplitude $\mathcal{A}_s$ and spectral index $n_s$ of primordial perturbations is quite simple to understand: changing $\mathcal{A}_s$ by a factor means multiplying all $C(l)$ by the same factor. Shifting $n_s \to n_s + \alpha$ changes the small-scale $C(l)$ by a factor $(l/l_p)^\alpha$, where $l_p$ is the angular wavenumber corresponding to the pivot scale $k_p$. This is not quite correct on large scales due to the wide support of $j_l$ for low $l$.

However, we also need to consider the optical depth due to *reionization*. After recombination, the gas in the universe was neutral. On the other hand, most of the gas we observe in the late universe is ionized; for example, we see no evidence for neutral gas in the absorption spectra of high-redshift quasars until we go back as far as $z \sim 6$ (Bouwens et al., 2015). So at some point the gas had to be *reionized*.[4] We currently believe that this happened between redshifts 15 and 6. After reionization, the CMB photons could scatter off the now-free electrons again. If enough scattering takes place, that is, if the optical depth $\tau_{rei} \equiv \tau(\eta_{late})$ back to some time $\eta_{late}$ after the end of the recombination epoch is high enough, isotropy is restored; equivalently, primordial anisotropies are washed out.

To study this quantitatively, imagine a photon traveling in our direction with temperature $T(1+\Theta)$, where $T$ is the background temperature and $\Theta$ is the perturbation. If these

[4] Unlike *recombination*, this name is actually accurate.

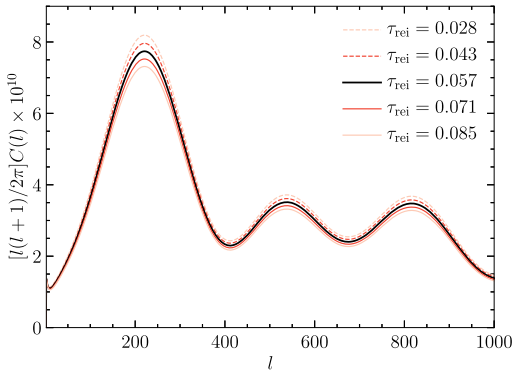



**FIGURE 9.16** Effect on the CMB power spectrum of varying the optical depth to reionization. On scales $l \gtrsim 150$, the effect is essentially an overall multiplicative factor, while the CMB is insensitive to $\tau_{\rm rei}$ on very large scales.

photons hit a region with optical depth $\tau_{\rm rei}$, only a fraction $e^{-\tau_{\rm rei}}$ will escape and continue on their way to us. In addition to these, we will observe a fraction $1 - e^{-\tau_{\rm rei}}$ scattered into the beam while traveling through the ionized region (since scattering conserves the total number of photons). These scattered-in photons come from all directions, so we can assume them to have the mean temperature $T$. So the temperature we see today is

$$T(1+\Theta)e^{-\tau_{\rm rei}} + T(1-e^{-\tau_{\rm rei}}) = T(1+\Theta\, e^{-\tau_{\rm rei}}). \tag{9.82}$$

Subtracting from this the mean temperature $T$ tells us that the fractional anisotropy will be the primordial one set up at $z \simeq 1100$ multiplied by $e^{-\tau_{\rm rei}}$. This scattering, however, affects only those perturbations within the horizon at the time of reionization, so only multipoles $l$ larger than $\eta_0/\eta_{\rm rei}$ will be suppressed by $e^{-\tau_{\rm rei}}$; small $l$ will be unaffected. This is seen in Fig. 9.16, which shows the effect of changing $\tau_{\rm rei}$. Clearly, increasing $\tau_{\rm rei}$ suppresses the anisotropies on small scales, but leaves them unchanged for $l \lesssim 100$.

This explains why we considered reionization together with the amplitude and spectral index: a change in $\mathcal{A}_s$, together with $n_s$, can largely mimic the effect of $\tau_{\rm rei}$, especially considering the fact that the cosmic variance error on the $C(l)$ is largest at low $l$. Conversely, the uncertain value of $\tau_{\rm rei}$ is by far the leading source of uncertainty on $\mathcal{A}_s$.

### 9.7.3 Baryon and CDM densities

The final variations we will consider are changes in the baryon density $\Omega_{\rm b}h^2$ and the CDM density $\Omega_{\rm c}h^2$. In each case, we keep a Euclidean universe and compensate the change in the density parameters through $\Omega_\Lambda$. As expected from our considerations at the beginning of this chapter, these changes lead to richer variations in the anisotropy spectrum than a mere shift and tilt; instead, they induce a small relative shift in the locations of the peaks and troughs in the spectrum, as well as changing their amplitudes. To understand these effects, it is important to recall that, since inhomogeneities on scales $k$ show up at $l = k\eta_0$ in a Euclidean universe, the peaks in a Euclidean universe will show up at $l_{\rm pk} \simeq k_{\rm pk}\eta_0 \simeq$

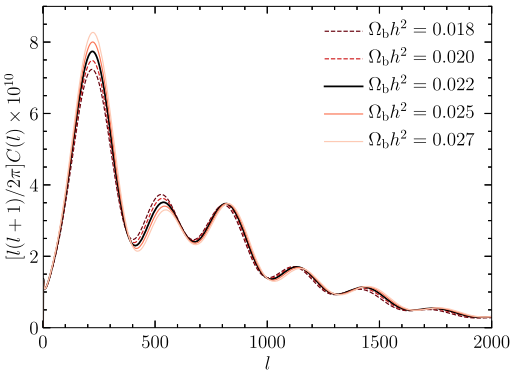



**FIGURE 9.17** Changes in the anisotropy spectrum as the baryon density $\Omega_b h^2$ is varied.

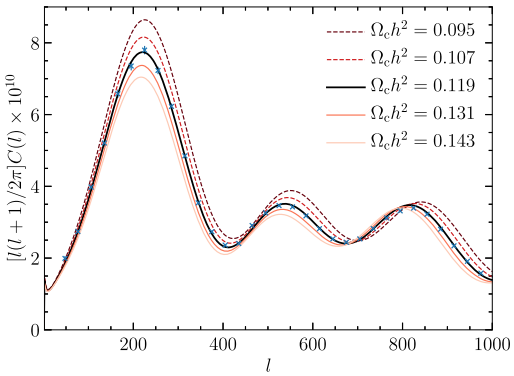



**FIGURE 9.18** Changes in the anisotropy spectrum as the CDM density $\Omega_c h^2$ is varied. Also shown are binned Planck measurements (Planck Collaboration, 2018b); the error bars are so small that they are only discernible for $l$ around and below the first peak. Clearly, $\Omega_c h^2$ and $\Omega_b h^2$ can be determined very precisely.

$n\pi\eta_0/r_s(\eta_*)$ (Eq. (9.27), but see the discussion in Sect. 9.6.2 that argues that the actual value of $l_{\rm pk}$ is $\sim 25\%$ lower).

The effects of changing the *baryon density* (Fig. 9.17) are a shift in the peak locations, due to the change in the sound horizon $r_s(\eta_*)$, as well as modifications in the heights of the peaks. We have already touched on the ways in which the anisotropy spectrum depends on the baryon density. The foremost, clearly visible in Fig. 9.17, is that the ratio of the heights of the odd to even peaks is higher when the baryon density is large. The second change due to $\Omega_b h^2$ is that an increased baryon density reduces the diffusion length (increases $k_D$). Therefore, a larger baryon density means damping moves to smaller angular scales, so the anisotropy spectrum on scales $l > 1000$ is larger in a high-$\Omega_b h^2$ model. This characteristic combination of effects allows for very tight constraints on $\Omega_b h^2$; the parameter variations around the fiducial values shown in Fig. 9.17 are ruled out by the data at high significance.

Next, we consider the effect of changing the *cold dark matter density* $\Omega_c h^2$ (Fig. 9.18). Part of the effect is changing the driving term for the acoustic oscillations (since the gravitational potential is dominated by CDM), which is similar to what a modification to the baryon density does. In addition, CDM determines to a large extent the epoch of equality, affecting both the evolution of perturbations (more growth for increased $\Omega_c h^2$) and the early ISW effect (less ISW for increased $\Omega_c h^2$, since the potentials decay less after recombination). The $C(l)$ are similarly sensitive to a given fractional change in $\Omega_c h^2$ as one in $\Omega_b h^2$. We also show binned power spectrum measurements—measurements for a range of $l$ combined into a single data point for presentation purposes—of the Planck team. They lie right on top of the fiducial Euclidean $\Lambda$CDM prediction. Given the barely visible error bars, the data precisely constrain $\Omega_c h^2$, ruling out at high significance the alternative parameter values we have shown for illustration.

## 9.8 Summary

The observed CMB anisotropies are a combination of three contributions (Eq. (9.59)):

- $\Theta_0 + \Psi$, which includes intrinsic photon temperature perturbation and gravitational redshift, and which we loosely refer to as "monopole." This contribution shows acoustic oscillations, whose behavior is to first order described by the semi-analytic tight-coupling solution.
- The Doppler-shift contribution $3\Theta_1$, which shows the same acoustic oscillations but is out of phase (as generally the case for velocities and positions of oscillators).
- The ISW contribution due to the time evolution of gravitational potentials around recombination and at late times. Unlike the first two, this is an integrated contribution along the line of sight, and has a smooth scale dependence.

The sum of these three contributions (and the associated cross-correlations) leads to the characteristic angular power spectrum of the CMB. The $C(l)$ contain rich information in particular on $\Omega_c h^2$ and $\Omega_b h^2$, but also on curvature and the amplitude of primordial perturbations $\mathcal{A}_s$ and the spectral index $n_s$ (although the constraint on $\mathcal{A}_s$ is limited by the degeneracy with the optical depth $\tau_{\rm rei}$). The CMB becomes even more powerful in combination with the large-scale structure probes discussed in Ch. 11. This combination has already led to percent-level constraints on the curvature and cosmological-constant parameters $\Omega_K$, $\Omega_\Lambda$.

Our semi-analytic approach to the CMB anisotropies follows Hu and Sugiyama (1995), whose extremely clear presentation is recommended as further reading. The benchmark for measurements of CMB temperature anisotropies is now set by the Planck satellite, at least for $l \lesssim 2000$, for which Planck's error bars have reached the fundamental cosmic variance limit. The latest data release is described in Planck Collaboration (2018a), while the parameter constraints, along with excellent concise descriptions of other data sets, are presented in Planck Collaboration (2018b).

## Exercises

**9.1** Most of this book is devoted to understanding adiabatic perturbations with the initial conditions derived in Ch. 7. Another class of perturbations are *isocurvature* perturbations with initial conditions $\Theta_0 = \Psi = \Phi = 0$. Physically, they correspond to relative perturbations between different particle species such that the total density perturbation is zero. Show that on large scales, these initial conditions imply that

$$\Theta_0(\boldsymbol{k}, \eta_*) + \Psi(\boldsymbol{k}, \eta_*) = 2\Psi(\boldsymbol{k}, \eta_*). \tag{9.83}$$

**9.2** The equation for a harmonic oscillator with a drag term is

$$m\ddot{x} + b\dot{x} + kx = 0. \tag{9.84}$$

Find the solutions to this equation if $k/m > (b/2m)^2$. What is the frequency of oscillations? How does this differ from the undamped ($b = 0$) solution? What is the other effect of nonzero $b$ besides the change in frequency?

**9.3** Determine $R(\eta_*)$ for the fiducial value of $\Omega_{\rm b}h^2$, and variations by $\pm 20\%$ around that value. Plot the sound speed as a function of scale factor for these three values of $\Omega_{\rm b}h^2$.

**9.4** Show that the sound horizon can be expressed in terms of the conformal time as

$$r_s(\eta) = \frac{2}{3k_{\rm eq}}\sqrt{\frac{6}{R(\eta_{\rm eq})}}\ln\left\{\frac{\sqrt{1+R} + \sqrt{R + R(\eta_{\rm eq})}}{1 + \sqrt{R(\eta_{\rm eq})}}\right\}, \tag{9.85}$$

where $k_{\rm eq}$ is given in Eq. (8.39).

**9.5** Obtain the WKB solution to Eq. (9.20). Write

$$\Theta_0 = Ae^{iB} \tag{9.86}$$

with $A$ and $B$ real. Show that the homogeneous part of Eq. (9.20) breaks up into two equations, coming from the real and imaginary parts:

$$\text{Real:}\quad -(B')^2 + \frac{A''}{A} + \frac{R'}{1+R}\frac{A'}{A} + k^2c_s^2 = 0, \tag{9.87}$$

$$\text{Imaginary:}\quad 2B'\frac{A'}{A} + B'' + \frac{R'}{1+R}B' = 0. \tag{9.88}$$

Find $B$ using the real part and the fact that $B$ changes much more rapidly than $A$. Then, use the imaginary equation to determine $A$. Show that the homogeneous solutions obtained in this way differ from the simple oscillatory solutions of Eq. (9.23) by a factor of $(1+R)^{1/4}$.

**9.6** Obtain a numerical solution for $\Theta_0 + \Psi$ and $\Theta_1$ at recombination by carrying out the integrals in Eqs. (9.26) and (9.28). To do this you will need expressions for the gravitational potentials. You can interpolate these from the output of CAMB or CLASS. Alternatively, use the fitting formulas given in Hu and Sugiyama (1995). Compare the result with the full numerical Boltzmann solution from CAMB or CLASS.

**9.7** Our treatment of diffusion damping neglected the effect of polarization. Go through the same expansion in $\tau'^{-1}$ that we carried out in Sect. 9.4, this time accounting for polarization. Show that this changes the factor of 8/9 in Eq. (9.42) to 16/15. This beautiful result was obtained by Zaldarriaga and Harari (1995).

**9.8** Assume that all electrons associated with hydrogen stay ionized and set $R = 0$. Evaluate the damping scale, $k_D$, defined in Eq. (9.42). Show that in this limit, the damping scale is given by Eq. (9.44), where

$$f_D(y) = 5\sqrt{1+1/y} - \frac{20}{3}(1+1/y)^{3/2} + \frac{8}{3}\left[(1+1/y)^{5/2} - 1/y^{5/2}\right]. \tag{9.89}$$

**9.9** Show that

$$\int d\Omega Y_{lm}(\hat{\boldsymbol{p}})\mathcal{P}_{l'}(\hat{\boldsymbol{p}}\cdot\hat{\boldsymbol{k}}) = \frac{4\pi}{2l+1}Y_{lm}(\hat{\boldsymbol{k}})\delta_{ll'}. \tag{9.90}$$

**9.10** There is a different way to go from the inhomogeneous temperature field at recombination, $\Theta_0(\boldsymbol{x}, \eta_*)$ or $\Theta_0(\boldsymbol{k}, \eta_*)$, to the anisotropy pattern today, $a_{lm}$, than that given in the text.

**(a)** Assume that the photons we see today from direction $\hat{\boldsymbol{p}}$ come from the surface of last scattering: $\Theta(\boldsymbol{x}_0, \hat{\boldsymbol{p}}, \eta_0) = (\Theta_0 + \Psi)(\boldsymbol{x} = \chi_*\hat{\boldsymbol{p}}, \eta_*)$ where $x_0$ is our position. That is, neglect dipole and ISW terms. Fourier transform the right-hand side and expand the left in terms of spherical harmonics to get

$$\sum_{lm} a_{lm}Y_{lm}(\hat{\boldsymbol{p}}) = \int \frac{d^3k}{(2\pi)^3}e^{i\boldsymbol{k}\cdot\hat{\boldsymbol{p}}\chi_*}(\Theta+\Psi)(\boldsymbol{k}, \eta_*). \tag{9.91}$$

Now expand the exponential using Eq. (C.17). Equate the coefficients of $Y_{lm}(\hat{\boldsymbol{p}})$ to get an expression for $a_{lm}$.

**(b)** Square the $a_{lm}$ you got in **(a)** and take the expectation value to get an expression for $C(l)$. You should recapture the expression in Eq. (9.74) when only including the monopole term in Eq. (9.59).

**9.11** Show that the cross-terms from the monopole and dipole are suppressed when summing over all modes. The monopole is proportional to $j_l(k\eta_0)$ while the dipole is proportional to $j_l'(k\eta_0)$. Compute the three possible integrals

$$\int_0^\infty dx j_l j_l; \qquad \int_0^\infty dx j_l j_l'; \qquad \int_0^\infty dx j_l' j_l'. \tag{9.92}$$

Show that the integrals of the squares ($j_l^2$ and $(j_l')^2$) are much larger than the integral of the cross-term $j_l j_l'$. Do the integrals for $l = 10$ up to $l = 200$.

**9.12** Determine the locations of the peaks and troughs in the CMB anisotropy spectrum if the universe is Euclidean with a cosmological constant. Keep the sound horizon fixed in this calculation by fixing $\Omega_{\rm m}h^2$ to the value in the fiducial cosmology. The peak positions then depend only on the distance to the last-scattering surface, $\eta_0 - \eta_*$. Consider two Euclidean models: $(i)$ $\Omega_\Lambda = 0$ (so that $\Omega_{\rm m} = 1$) and $(ii)$ $\Omega_\Lambda = 0.7$ (so that $\Omega_{\rm m} = 0.3$). What value of $h$ is needed in the two cases to keep $\Omega_{\rm m}h^2$ fixed? Determine $\eta_0 - \eta_*$ in each case (in the cosmological constant case, you will have to do the integral numerically). Compare your result with Fig. 9.17.

**9.13** Compute the distance to the last-scattering surface in a Euclidean model with dark energy described by $\Omega_{\rm de} = 0.7$ (today) and $w = -0.5$. Compare the expected peak locations in the anisotropy spectrum with the cosmological constant model of the previous exercise.

**9.14** Derive the effects of reionization using the Boltzmann equation. Neglect the gravitational potentials, the velocity, and $\Theta_0$ in the Boltzmann equation for photons. Start with a spectrum $\Theta_l(\eta_{\rm in})$ at an initial time $\eta_{\rm in}$ after recombination and evolve until today. Show that the moments are indeed suppressed by $e^{-\tau_{\rm rei}}$.

**9.15** Assume that recombination took place instantaneously. Show that the solution for the $l$th moment due to tensor perturbations (Eq. (6.86)) is

$$\Theta_{l,t}^{\rm T}(k,\eta_0) = -\frac{1}{2}\int_{\eta_*}^{\eta_0} d\eta\, \left(h_t^{\rm TT}\right)' j_l[k(\eta_0-\eta)], \tag{9.93}$$

for either tensor-mode polarization $t = +, \times$.

**9.16** Using the decomposition for tensor modes given in Eq. (6.85), find the contribution to the $C(l)$ from $\Theta_l^{\rm T}(k,\eta_0)$. That is, show that the analogue of Eq. (9.74) for tensors is

$$C^{\rm T}(l;t) = \frac{(l-1)l(l+1)(l+2)}{\pi}\int_0^\infty dk\, k^2 \times \left|\frac{\Theta_{l-2,t}^{\rm T}}{(2l-1)(2l+1)} + 2\frac{\Theta_{l,t}^{\rm T}}{(2l-1)(2l+3)} + \frac{\Theta_{l+2,t}^{\rm T}}{(2l+1)(2l+3)}\right|^2, \tag{9.94}$$

where $t = +, \times$ denotes the two tensor-mode polarizations.

**9.17** Determine the spectrum of the temperature anisotropies due to gravitational waves produced during inflation.

**(a)** Combine the results of the previous two exercises, your solution to Exercise 6.12, and the primordial amplitude of gravitational waves in Eq. (7.102) to find the large-angle angular power spectrum $C^{\rm T}(l)$ due to primordial gravitational waves.

**(b)** Tensor anisotropies can be parametrized empirically by

$$r_2 \equiv \frac{C^{\mathrm{T}}(l=2)}{C(l=2)}, \tag{9.95}$$

i.e. the ratio of the contributions to the temperature quadrupole from tensor and scalar perturbations. We already derived an expression for the scalar $C(l=2)$ in Eq. (9.80). Find $C^{\mathrm{T}}(l=2)$ and compute $r_2$. You can set $n_{\mathrm{T}}=0$ for this exercise. Compare with the conventional definition of the tensor-to-scalar ratio $r$ in Sect. 7.6.

# 10

# The polarized CMB

Our treatment of the CMB so far has been focused on the temperature anisotropies. However, Compton scattering before decoupling also induces polarization anisotropies. Polarization opens up a new dimension in the study of the CMB. On the one hand, it in principle doubles the amount of information in the CMB about the scalar perturbations we have focused on so far. As we will see in this chapter, however, the promise of polarization goes well beyond this: gravitational waves—tensor perturbations—produce a particular pattern of polarization that cannot be mimicked by scalar perturbations on large scales. Therefore, polarization offers a unique way of searching for gravitational waves produced during inflation.

Most of the physics involved in CMB polarization is already familiar to us. However, the treatment is somewhat more difficult technically for two reasons. One difficulty is in describing a polarization pattern on the sky, which is no longer a simple function of the position on the sky as the temperature is. Instead, we are dealing with a "headless vector." We will simplify this obstacle by assuming the flat-sky approximation (Sect. 10.1). This means that our results will only be valid on small angular scales, but this is a reasonable price to pay for the much simpler derivation that this approximation offers; we will not miss any of the essential physics. The other difficulty is keeping track of the geometry of Compton scattering of polarized radiation. We will do this in two steps in Sect. 10.2 and Sect. 10.3. This geometry is crucial in order to understand why polarization is such a powerful probe of tensor modes.

The results of Sect. 10.1 will further become very useful when dealing with galaxy ellipticities and gravitational lensing in Ch. 13. In fact, we will be able to take over all results from this section essentially one-to-one.

## 10.1 Polarization

The polarization of a radiation field is measured most simply by inserting a polarizer in front of the detector, which allows only waves oscillating in a particular direction (in the plane perpendicular to propagation) to pass through (Fig. 10.1). By plotting the intensity recorded by the detector as a function of the orientation of the polarizer, one can measure the polarization. If we rotate the polarizer by 180°, we obtain the same result: polarization is a *headless vector*. Then, if we denote the unit direction vector of the polarizer with $\hat{\boldsymbol{m}}$ (where $\hat{\boldsymbol{m}} \cdot \hat{\boldsymbol{p}} = 0$, with $\hat{\boldsymbol{p}}$ being the wavevector of the radiation), the flux of radiation incident on the detector cannot depend on the sign of $\hat{\boldsymbol{m}}$. Therefore, it must be a quadratic function

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00016-4
Copyright © 2021 Elsevier Inc. All rights reserved.

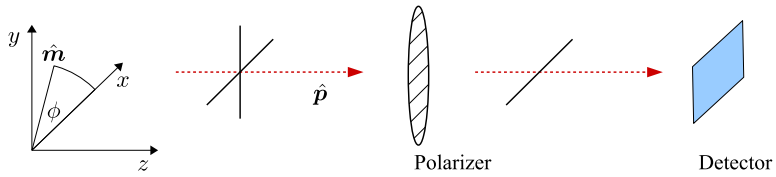



**FIGURE 10.1** Measuring polarized light. Photons coming in along the $\hat{\boldsymbol{p}}$-direction (here, $\hat{\boldsymbol{p}} = \hat{\boldsymbol{e}}_z$) are polarized in the plane perpendicular to $\hat{\boldsymbol{p}}$. Polarization is measured by sending the light through a polarizer, which allows only light with a certain linear polarization to pass through. By rotating the polarizer around the $\hat{\boldsymbol{p}}$-axis, the degree and direction of the polarization of light can be measured. It is a headless vector, and can be described by its azimuthal angle $\phi$ with respect to the $x$-axis in the $x$–$y$ plane perpendicular to $\hat{\boldsymbol{p}}$.

of $\hat{\boldsymbol{m}}$. We write

$$I_{\rm det}(\hat{\boldsymbol{m}}) = I_{ij}\hat{\boldsymbol{m}}^i\hat{\boldsymbol{m}}^j, \tag{10.1}$$

where $I_{ij}$ describes the polarization of radiation (*polarization tensor*). Notice that this is a $2\times 2$ matrix, since $\hat{\boldsymbol{m}}$ is constrained to lie in the plane perpendicular to $\hat{\boldsymbol{p}}$. Clearly, $I_{ij}$ can be taken to be symmetric.[1] Moreover, for unpolarized light, the detected intensity $I_{\rm det}(\hat{\boldsymbol{m}})$ is identical in the $\hat{\boldsymbol{m}}_x$- and $\hat{\boldsymbol{m}}_y$-directions, so $I_{ij}\propto\delta_{ij}$. Let us then write

$$I_{ij} = \begin{pmatrix} I+Q & U \\ U & I-Q \end{pmatrix}. \tag{10.2}$$

The diagonal elements $I$ are the intensity, which is what we studied in Ch. 9 (as temperature $T$, with a uniform part and a perturbation $\Theta$). The two new variables $Q$ and $U$ describe polarization, and they are illustrated in Fig. 10.2. Students of electricity and magnetism might recognize $I$, $Q$, and $U$ as three of the four Stokes parameters used to describe polarization. The fourth, $V$, is absent since we ignore circular polarization. So, instead of dealing with a single distribution function $f(\boldsymbol{x},\hat{\boldsymbol{p}},\eta)$, we now have two additional distribution functions $f_Q$ and $f_U$ to consider. In fact, we will see that a slightly different parametrization is more useful than $Q$ and $U$.

Our goal in this chapter will be to derive the statistics of $Q$ and $U$ that we expect in the CMB. For this, we will turn again to the solution of the Boltzmann equation derived in Ch. 9. However, before doing that, we have to deal with some subtleties related to the nature of polarization. The statistics of the temperature were straightforward to derive because the temperature is a coordinate-invariant quantity. This is not the case for polarization: if we measure the polarization along the $x$-axis, say, corresponding to $I_{xx}$, and then rotate our coordinate system, the value of $I_{xx}$ changes. $Q$ and $U$ change accordingly as well. Fortunately, we already have developed the tools to deal with this in Sect. 6.1; we now simply have to apply them to two instead of three spatial dimensions.

[1] In general, $I_{ij}$ is Hermitian; throughout we will neglect circular polarization since it is not generated by cosmological perturbations, in which case $I_{ij}$ can be taken to be real and symmetric.

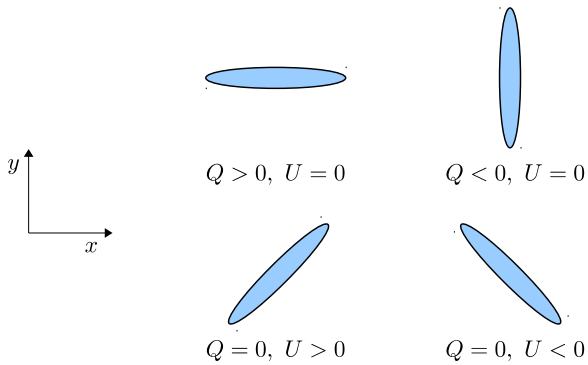



**FIGURE 10.2** Definition of polarization components $Q$ and $U$ in the plane perpendicular to the incoming light. Unpolarized light has $Q = U = 0$.

In this chapter, we will work in the flat-sky approximation. This significantly simplifies the derivation, while keeping all of the physics untouched. Hence, we treat the position on the sky $\boldsymbol{\theta}$ as a 2D vector on the $x$–$y$ plane. Correspondingly, instead of considering multipole moments $l$, $m$, we define the 2D vector $\boldsymbol{l}$ as the Fourier counterpart of $\boldsymbol{\theta}$. For example, the temperature, proportional to the intensity $I$, in Fourier space becomes

$$T(\boldsymbol{l}) = \int d^2\theta\, T(\boldsymbol{\theta}) e^{i\boldsymbol{l}\cdot\boldsymbol{\theta}}. \tag{10.3}$$

Now, we can write Eq. (10.2) as

$$I_{ij} = I\,\delta_{ij} + I^{\mathrm{T}}_{ij}, \tag{10.4}$$

where $I^{\mathrm{T}}_{ij}$ is traceless and contains the information needed about the two polarization states. Instead of $Q$ and $U$, though, we want to characterize those two states by their behavior under rotations. Just as we decomposed the elements of the spatial metric perturbation $h_{ij}$ in Sect. 6.1 into scalar, vector, and tensor perturbations, here we want to do the same with $I^{\mathrm{T}}_{ij}$. It is actually a bit simpler in two dimensions: $I_{ij}$ has three independent components, and so the traceless part $I^{\mathrm{T}}_{ij}$ has only two. One of these is a scalar which we can extract by taking the combination $l^i l^j I^{\mathrm{T}}_{ij}/l^2$. We will call this $E(\boldsymbol{l})$. The other, which we will denote as $I^{\mathrm{TT}}_{ij}(\boldsymbol{l})$, is a transverse-traceless tensor (so $l^i I^{\mathrm{TT}}_{ij}(\boldsymbol{l}) = 0$). You can check that the following decomposition is consistent with what we described:

$$I^{\mathrm{T}}_{ij}(\boldsymbol{l}) = 2\left(\frac{l_i l_j}{l^2} - \frac{1}{2}\delta_{ij}\right) E(\boldsymbol{l}) + I^{\mathrm{TT}}_{ij}(\boldsymbol{l}). \tag{10.5}$$

We will soon see that the scalar component $E$ couples to both scalar and tensor metric perturbations, while the transverse-traceless component couples *only* to tensor metric perturbations. This is one more reason why this decomposition is powerful. First, though,

let us connect these scalar and tensor pieces to the original $Q/U$ decomposition. We extract the scalar component from $I^{\rm T}_{ij}$ by contracting with $l^i l^j/l^2$:

$$\begin{aligned} E(\boldsymbol{l}) &= \frac{l^i l^j}{l^2} I^{\rm T}_{ij} \\ &= (\cos^2\phi_l - \sin^2\phi_l)\, Q(\boldsymbol{l}) + 2\sin\phi_l \cos\phi_l\, U(\boldsymbol{l}), \end{aligned} \tag{10.6}$$

where in the second line we have written out in components by introducing the azimuthal angle $\phi_l$ of the 2D wavevector: $\boldsymbol{l} = (l_x, l_y) = (\cos\phi_l, \sin\phi_l)\, l$. Using the trigonometric addition formulas, we can simplify this to

$$E(\boldsymbol{l}) = \cos 2\phi_l\, Q(\boldsymbol{l}) + \sin 2\phi_l\, U(\boldsymbol{l}). \tag{10.7}$$

Next, we can use Eq. (10.5) for $I^{\rm TT}$. First,

$$\begin{aligned} I^{\rm TT}_{12}(\boldsymbol{l}) &= I^{\rm T}_{12} - 2\frac{l_1 l_2}{l^2} E(\boldsymbol{l}) \\ &= U(\boldsymbol{l}) - \sin 2\phi_l\, (\cos 2\phi_l\, Q(\boldsymbol{l}) + \sin 2\phi_l\, U(\boldsymbol{l})) \\ &= (1 - \sin^2 2\phi_l) U(\boldsymbol{l}) - \sin 2\phi_l \cos 2\phi_l\, Q(\boldsymbol{l}) \\ &= \cos 2\phi_l B(\boldsymbol{l}), \end{aligned} \tag{10.8}$$

where we have again used the trigonometric formulae and defined

$$B(\boldsymbol{l}) = -\sin 2\phi_l\, Q(\boldsymbol{l}) + \cos 2\phi_l\, U(\boldsymbol{l}). \tag{10.9}$$

You can easily show, via a very similar calculation, that the other remaining elements of $I^{\rm TT}_{ij}$ are also given in terms of $B$:

$$\frac{1}{2}(I^{\rm TT}_{11} - I^{\rm TT}_{22})(\boldsymbol{l}) = -\sin 2\phi_l\, B(\boldsymbol{l}). \tag{10.10}$$

So we have decomposed the polarization tensor $I^{\rm T}_{ij}$, which has two independent components $Q$ and $U$, into the scalar part $E(\boldsymbol{l})$, and $B(\boldsymbol{l})$, which describes the tensor part. This leads to our final decomposition of the polarization tensor:

$$I^{\rm T}_{ij}(\boldsymbol{l}) = \begin{pmatrix} \cos 2\phi_l & \sin 2\phi_l \\ \sin 2\phi_l & -\cos 2\phi_l \end{pmatrix} E(\boldsymbol{l}) + \begin{pmatrix} -\sin 2\phi_l & \cos 2\phi_l \\ \cos 2\phi_l & \sin 2\phi_l \end{pmatrix} B(\boldsymbol{l}). \tag{10.11}$$

As an example, let us specialize to a single mode with a wavevector $\boldsymbol{l} = l_0 \hat{\boldsymbol{e}}_x$ along the $x$-axis (i.e. $\phi_l = 0$). Then, the polarization pattern in real space becomes

$$I^{\rm T}_{ij}(\boldsymbol{\theta}) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} e^{il_0\theta_x} E_0 + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} e^{il_0\theta_x} B_0. \tag{10.12}$$

The result is shown in Fig. 10.3. The $E$-mode varies in strength in the same direction as, or perpendicular to, its orientation. This conjures images of an electric field. An electric field

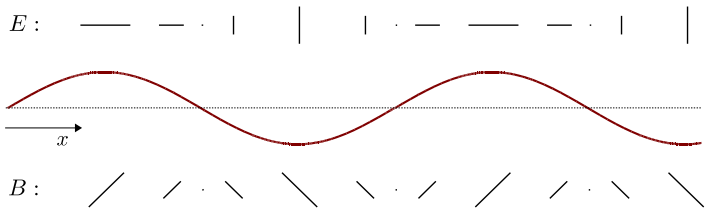



**FIGURE 10.3** Polarization generated by a single plane-wave perturbation along the $x$-axis (i.e. $\boldsymbol{k} = k\hat{\boldsymbol{e}}_x$), for an $E$-mode (top) and a $B$-mode (bottom). The length of each headless arrow indicates the strength of polarization. These patterns are distinguishable by looking at how the polarization angle aligns with the wavevector of the mode (the direction in which the polarization changes). Notice that here and in Fig. 10.4 we show the polarization pattern for photons propagating out of the page (along the $z$-axis).

from a point source, $\boldsymbol{E} = q\hat{\boldsymbol{r}}/r^2$, varies in strength as one moves away from the point source, but always points in the same direction: radially away from the source. As one moves in the direction of the field, the strength of the field decreases. The $B$-mode, on the other hand, varies in strength in a different direction from that in which it is pointing (namely by 45°), just like a magnetic field. Notice that one can generate the $B$-mode pattern from the $E$-mode pattern by rotating each polarization direction by 45°.

Motivated by this analogy, let us consider a superposition of plane waves in the $x$–$y$ plane that have equal wavelength and phase at the origin, as well as amplitude. That is, they only differ by the azimuthal angle $\phi_l$. The resulting polarization patterns (try to work them out!) for $E$- and $B$-modes look like those shown in Fig. 10.4. These patterns are clearly distinct, with the $B$-modes showing the characteristic "swirly" form. One can also glean from this figure that $E$- and $B$-modes have different parity: $E$ is parity even, i.e. it stays the same if we imagine flipping the page, while $B$ is parity-odd, i.e. it changes sign. It comes as no surprise then that we can distinguish the $E$- and $B$-components in an observed polarization map.

Beyond pretty illustrations, we have accomplished an important goal: $E(\boldsymbol{l})$ is now a scalar on the sky just like the temperature (while $B(\boldsymbol{l})$ is a pseudoscalar; you can show this, and the fact that the same does not hold for $Q$ and $U$, in Exercise 10.1). This means we can compute the power spectra straightforwardly. A further significance of Eq. (10.11) is that we can use a convenient coordinate system to tell whether a given physical process generates the $E$ or $B$ components. We see from Eq. (10.12) that a mode with a wavevector $\boldsymbol{l}$ along the $x$-axis which only has a $Q$ component, and no $U$, corresponds to a pure $E$-mode, a conclusion which then holds for any orientation of the wavevector.

## 10.2 Generating polarization from Compton scattering

Light traveling in the $z$-direction corresponds to electric and magnetic fields oscillating in the $x$–$y$ plane, i.e., transverse to the direction of propagation. If the intensity along the two transverse directions is equal, then the light is unpolarized. Until now, when we have considered the CMB, we have been implicitly studying this case. Now we must account

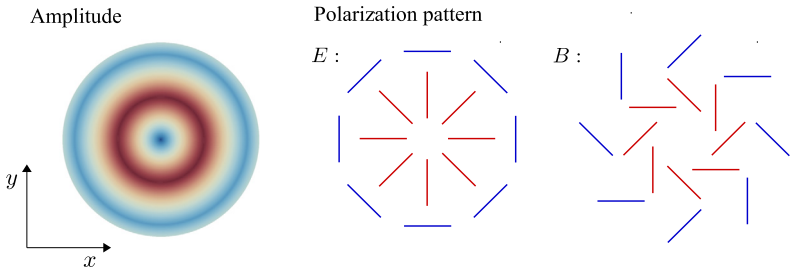



**FIGURE 10.4** Similar to Fig. 10.3, but now polarization generated by a radial wave in the $x$–$y$ plane is shown (left panel), i.e. a superposition of plane waves with equal phase and amplitude but different azimuthal angles $\phi_l$. This configuration illustrates the difference between $E$- and $B$-mode patterns very clearly (middle and right panels, respectively; again for light coming out of the page). In each case, we show a peak (red (light gray in print version), inner set of lines) and a trough (blue (dark gray in print version), outer set of lines) of the radial wave.

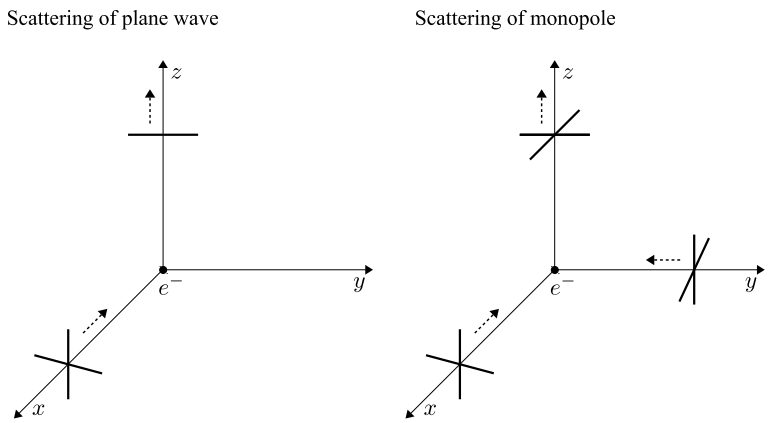



**FIGURE 10.5** *Left panel:* unpolarized plane-wave radiation moving toward the origin along the $x$-axis is scattered by an electron into the $+z$-direction. Only the $y$ component of the radiation remains after scattering. Since there was no incoming $x$ polarization, the outgoing radiation is polarized in the $y$-direction. *Right panel:* incoming isotropic (monopole) radiation produces no polarization. Here, since the incoming amplitudes from the $x$- and $y$-directions are equal, the outgoing intensities along both of these directions are equal, leading to unpolarized radiation (see also Hu and White, 1997).

for the possibility that the intensities in the two transverse directions are unequal: that the radiation is polarized.

Compton scattering is able to produce polarized radiation.[2] To see this, consider the left panel of Fig. 10.5 which shows a ray incident from the $+x$-direction. This (unpolarized) ray has equal intensity in the $y$- and $z$-directions. It scatters off an electron at the origin and gets deflected into the $+z$-direction (our line of sight). Since the outgoing direction is along the $z$-axis, none of the (incoming) intensity along the $z$-axis gets transmitted. By

[2] As in Ch. 5, we are only considering elastic scattering of photons off electrons here, which is a special case of Compton scattering known as Thomson scattering.

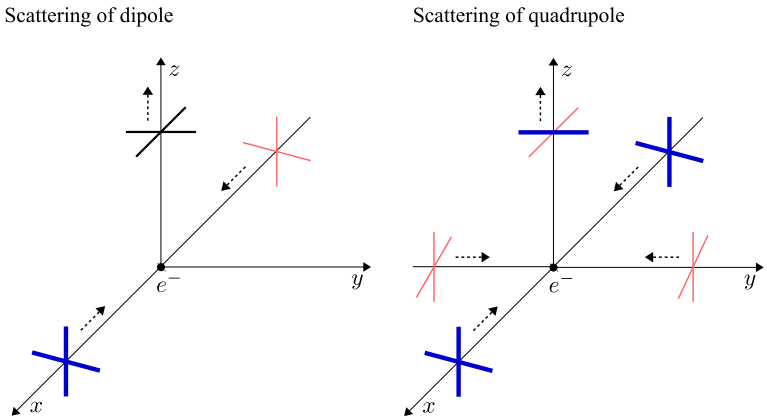



**FIGURE 10.6** *Left panel:* incoming radiation with a dipolar pattern produces no polarization. Heavy blue (thin red) lines denote hotter (colder) radiation. In this case, the incoming radiation is hotter than average (average is shown as medium black lines) from the $+x$-direction, and colder than average from the $-x$-direction (the incoming radiation along the $\pm y$-direction is of average temperature, as in the right panel of Fig. 10.5). The two rays from the $\pm x$-directions therefore produce the average intensity for the outgoing ray along the $y$-direction. The outgoing intensity along the $x$-direction is produced by the ray incident from the $\pm y$-directions. Since these have the average intensity, the outgoing intensity is also the average along the $x$-direction. The net result is outgoing unpolarized light. *Right panel:* incoming radiation with a quadrupolar pattern produces outgoing polarized light. The outgoing radiation has greater intensity along the $y$-axis than in the $x$-direction. This is a direct result of the hotter radiation incident from the $\pm x$-direction as compared to that from the $\pm y$-direction.

contrast, all of the intensity along the $y$-axis (which is perpendicular to both the incoming and the outgoing directions) is transmitted. The net result is polarized outgoing emission: the difference between the outgoing intensities in the two directions perpendicular to $\hat{\boldsymbol{e}}_z$ ($y$: fully transmitted; $x$: zero) is maximal.

Obviously, we cannot content ourselves with studying one incoming ray; we must generalize to radiation incident on an electron from all directions. When we do so, we begin to realize that producing polarization will not be quite as easy as it appears from the left panel of Fig. 10.5. Consider then the right panel, which shows a caricature of a much more relevant case: isotropic radiation incident on the electron from all directions. We say "caricature" because we have shown incoming rays from only two directions, the $+x$- and $+y$-directions; this will be sufficient for our argument. The $I_{xx}$ component of the outgoing ray's intensity comes from the radiation incident from the $y$ direction, while the outgoing $I_{yy}$ comes from the radiation incident from the $x$-axis. Since the incoming radiation from both directions has equal intensity (isotropic radiation), though, the outgoing wave has equal intensity along both axes, $I_{xx} = I_{yy}$: it is unpolarized.

Can anisotropic radiation produce polarization? The simplest example of anisotropy is a dipole pattern, a caricature of which is shown in the left panel of Fig. 10.6. Now the outgoing intensity along the $x$-axis comes from the $\pm y$-incident radiation, which has the average temperature (not drawn in the figure). The outgoing intensity along the $y$-axis is also average, since it is due to a superposition of a cold spot (the $-x$-direction) and a hot

spot (the $+x$-direction). The dipole pattern therefore only leads to unpolarized outgoing radiation.

To produce polarized radiation, the incoming radiation must have a nonzero quadrupole. This is shown in the right panel of Fig. 10.6. The hotter (colder) radiation incident from the $x$- ($y$-) direction produces higher (lower) intensity along the $y$- ($x$-) axis for the outgoing wave. Therefore, the intensity of the outgoing wave is greater along the $y$-axis than along the $x$-axis: the outgoing radiation is polarized. Fig. 10.6 depicts polarization in the $x$–$y$ plane, preferentially in the $y$-direction: from Fig. 10.2, we see that this pattern corresponds to $Q < 0$ and $U = 0$. Alternatively, had the incoming rays been rotated by 45° in the $x$–$y$ plane, the outgoing polarization would have been along the axis 45° from the $x$- and $y$-axes, and correspondingly produced a $U$ component.

The fact that Compton scattering produces polarization only when the incident field has a quadrupole moment has important ramifications for cosmology. Because electrons and photons are tightly coupled before recombination, the radiation field has a very small quadrupole. Therefore, we expect CMB polarization to be smaller than the temperature anisotropies.

## 10.3 Polarization from a single plane wave

The pictures of the previous subsection are important to gain a qualitative understanding of how Compton scattering produces polarization in the CMB. In order to study the phenomenon quantitatively, we must solve the Boltzmann equation, taking into account the two photon polarization states and the polarization dependence of Compton scattering in the collision term. We begin by considering a single plane-wave perturbation to the photon distribution.

We first need to define the polarization axes in the most general case when the incoming photon arrives from direction $\hat{\boldsymbol{n}}'$ (Fig. 10.7). When that direction was $\hat{\boldsymbol{e}}_x$, as in the previous section, it was clear that polarization was defined as the difference in the intensity along the two perpendicular directions, $y$ and $z$. Now, we denote the two axes perpendicular to this direction, describing the plane in which incoming rays are polarized, with $\hat{\boldsymbol{\epsilon}}'_1$ and $\hat{\boldsymbol{\epsilon}}'_2$. We continue to focus on outgoing photons in the $z$-direction, so we can choose the two outgoing polarization axes as $\hat{\boldsymbol{\epsilon}}_1 = \hat{\boldsymbol{e}}_x$ and $\hat{\boldsymbol{\epsilon}}_2 = \hat{\boldsymbol{e}}_y$. In short, the incoming polarization vectors are $\hat{\boldsymbol{\epsilon}}'_i$, the outgoing are $\hat{\boldsymbol{\epsilon}}_i$.

The polarization dependence of Compton scattering that we discussed in the previous section is neatly encapsulated by an additional factor in the amplitude squared for outgoing photons polarized in the $\hat{\boldsymbol{\epsilon}}_i$-direction. Summing over the two (irrelevant) electron spins as well as the *incoming* photon polarization, the amplitude squared in Eq. (5.18), now for the outgoing polarizations $i = 1, 2$ separately, receives an additional factor,

$$\sum_{3\text{ spins}} |\mathcal{M}|^2 \propto \sum_{j=1}^{2} |\hat{\boldsymbol{\epsilon}}_i(\hat{\boldsymbol{n}}) \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2. \tag{10.13}$$

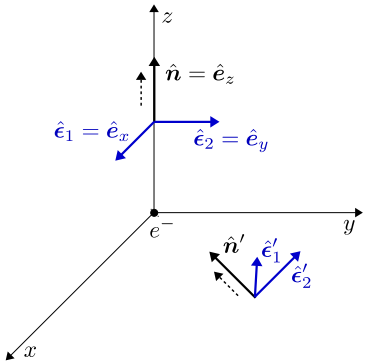



**FIGURE 10.7** Incoming photon from direction $\hat{\boldsymbol{n}}'$ scatters off an electron at the origin producing an outgoing photon in the direction $\hat{\boldsymbol{n}} = \hat{\boldsymbol{e}}_z$. The plane perpendicular to the incoming direction is spanned by the two polarization vectors, $\hat{\boldsymbol{\epsilon}}'_1 = \hat{\boldsymbol{e}}'_\theta$ and $\hat{\boldsymbol{\epsilon}}'_2 = \hat{\boldsymbol{e}}'_\phi$. The outgoing photon is in the $\hat{\boldsymbol{e}}_z$-direction, so the polarization vectors are $\hat{\boldsymbol{\epsilon}}_1 = \hat{\boldsymbol{e}}_x$ and $\hat{\boldsymbol{\epsilon}}_2 = \hat{\boldsymbol{e}}_y$.

Let us first calculate the $Q$ polarization which, from Eq. (10.2), is sourced by the difference between this cross-section for $i = 1$ and $i = 2$, i.e., the difference between the field strength in the $\hat{\boldsymbol{e}}_x$- and $\hat{\boldsymbol{e}}_y$-directions:

$$\sum_{j=1}^{2} |\hat{\boldsymbol{\epsilon}}_1(\hat{\boldsymbol{n}}) \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 - \sum_{j=1}^{2} |\hat{\boldsymbol{\epsilon}}_2(\hat{\boldsymbol{n}}) \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 = \sum_{j=1}^{2} \left( |\hat{\boldsymbol{e}}_x \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 - |\hat{\boldsymbol{e}}_y \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 \right). \tag{10.14}$$

Integrating over all incoming $\hat{\boldsymbol{n}}'$-directions leads to

$$Q(\hat{\boldsymbol{e}}_z) = A \int d\Omega' f(\hat{\boldsymbol{n}}') \sum_{j=1}^{2} \left( |\hat{\boldsymbol{e}}_x \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 - |\hat{\boldsymbol{e}}_y \cdot \hat{\boldsymbol{\epsilon}}'_j(\hat{\boldsymbol{n}}')|^2 \right). \tag{10.15}$$

Here $A$ is a normalization constant which will concern us in the next section, and $f(\hat{\boldsymbol{n}}')$ is the intensity of the radiation incoming from the $\hat{\boldsymbol{n}}'$-direction, where we integrate over all such directions. Note that $f$ depends only on $\hat{\boldsymbol{n}}'$, but not on $j$: this corresponds to the assumption that the incident radiation is unpolarized.

To take the dot products in Eq. (10.15), we need $\hat{\boldsymbol{\epsilon}}'_1$ and $\hat{\boldsymbol{\epsilon}}'_2$ in terms of their Cartesian coordinates. These are defined to be orthogonal to $\hat{\boldsymbol{n}}'$, whose components are

$$\hat{\boldsymbol{n}}' = (\sin\theta' \cos\phi',\ \sin\theta' \sin\phi',\ \cos\theta'). \tag{10.16}$$

We choose $\hat{\boldsymbol{\epsilon}}'_2$ to lie in the $x$–$y$ plane, which leads to

$$\hat{\boldsymbol{\epsilon}}'_2(\theta', \phi') = (-\sin\phi', \cos\phi', 0). \tag{10.17}$$

Then, $\hat{\boldsymbol{\epsilon}}'_1$ is completely specified by taking the vector product of $\hat{\boldsymbol{n}}'$ and $\hat{\boldsymbol{\epsilon}}'_2$:

$$\hat{\boldsymbol{\epsilon}}'_1(\theta', \phi') = (\cos\theta' \cos\phi', \cos\theta' \sin\phi', -\sin\theta'). \tag{10.18}$$

Alternatively, we can notice that $\hat{\boldsymbol{\epsilon}}'_1$ and $\hat{\boldsymbol{\epsilon}}'_2$ are given by the spherical coordinate vectors $\hat{\boldsymbol{e}}'_\theta$ and $\hat{\boldsymbol{e}}'_\phi$, respectively. Now the dot products become straightforward, and we find

$$\begin{aligned} Q(\hat{\boldsymbol{e}}_z) &= A\int d\Omega' f(\hat{\boldsymbol{n}}')\left[\cos^2\theta'\cos^2\phi' + \sin^2\phi' - \cos^2\theta'\sin^2\phi' - \cos^2\phi'\right] \\ &= -A\int d\Omega' f(\hat{\boldsymbol{n}}')\sin^2\theta'\cos 2\phi'. \end{aligned} \tag{10.19}$$

The angular dependence is proportional to the sum of the spherical harmonics $Y_{2,2} + Y_{2,-2}$ (Eq. (C.10)). Since the spherical harmonics are orthogonal, the integral will pick out the $l = 2$, $m = \pm 2$ components of the distribution $f$. That is, nonzero $Q$ will be produced if and only if the incident radiation has a quadrupole moment. This verifies the argument-by-pictures given in the previous section. It is straightforward to derive the corresponding expression for the $U$-component of polarization (Exercise 10.3),

$$U(\hat{\boldsymbol{e}}_z) = -A\int d\Omega' f(\hat{\boldsymbol{n}}')\sin^2\theta'\sin(2\phi'). \tag{10.20}$$

Now the angular dependence is proportional to $Y_{2,2} - Y_{2,-2}$. Again, only an incident quadrupole produces $U$ polarization.

We can now relate the outgoing $Q$ and $U$ fields to the moments of the incident unpolarized distribution. We will do this in several steps, starting with the $k$-vector along the $x$-axis, then generalizing this to the $x$–$z$ plane, and finally obtaining the result for an arbitrary direction of $\boldsymbol{k}$. The first step, however, already yields the main features of our final result.

The reason that we need to move step by step is that the direction of the quadrupole in the photon distribution $f(\hat{\boldsymbol{n}}')$ is determined by the direction of the wavevector. Recall that, in Ch. 5 we wrote the photon distribution as the sum of a zeroth-order piece—the uniform Planck distribution—and a perturbation, characterized by $\Theta(k,\mu)$ where $\mu$ is the dot product of the wave vector $\hat{\boldsymbol{k}}$ and the direction of the photon momentum. Thus, we now need to keep track of three directions: the wavevector $\hat{\boldsymbol{k}}$; the incoming direction $\hat{\boldsymbol{n}}'$; and the outgoing photon direction $\hat{\boldsymbol{n}}$. We already have used up some of our coordinate freedom by choosing $\hat{\boldsymbol{n}} = \hat{\boldsymbol{e}}_z$. In our case, the variable $\mu$ becomes $\mu = \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{n}}'$. Thus, $f(\hat{\boldsymbol{n}}')$ in Eq. (10.19) will be an expansion in Legendre polynomials with argument $\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{n}}'$, which is *not* equal to the cosine of $\theta'$, since $\theta'$ is the angle between the $z$-axis and $\hat{\boldsymbol{n}}'$. Relating $\mu$ to $\theta'$ and $\phi'$ therefore is not trivial, and we will proceed slowly.

Let us first consider the wave vector $\boldsymbol{k}$ to lie in the $x$-direction. Then,

$$\begin{aligned} \mu \equiv \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{n}}' &= \left(\hat{\boldsymbol{n}}'\right)_x \\ &= \sin\theta'\cos\phi' \end{aligned} \tag{10.21}$$

where the second equality follows from Eq. (10.16). Recall that we decomposed the perturbation $\Theta$ into a sum over Legendre polynomials, so

$$\Theta(k, \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}') = \sum_l (-i)^l (2l+1)\Theta_l(k)\mathcal{P}_l\left(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}'\right)$$
$$\to -5\Theta_2(k)\mathcal{P}_2(\sin\theta' \cos\phi'), \tag{10.22}$$

where the last line follows by considering only the relevant quadrupole part of the sum, and substituting our expression for $\mu$ (Eq. (10.21)). Notice also that, as we did in Ch. 8 and Ch. 9, we continue to focus on scalar perturbations in this section.

A plane wave with wavevector $\boldsymbol{k}$ pointing in the $x$-direction therefore produces

$$Q(\hat{\boldsymbol{e}}_z, \boldsymbol{k} \| \hat{\boldsymbol{e}}_x) = 5A\Theta_2(k) \int_0^\pi d\theta' \sin\theta' \int_0^{2\pi} d\phi' \mathcal{P}_2(\sin\theta' \cos\phi') \sin^2\theta' \cos 2\phi'. \tag{10.23}$$

Recall that $\mathcal{P}_2(\mu) = (3\mu^2 - 1)/2$. The $-1/2$ part of this gives no contribution to the integral since the $\phi'$ integral over $\cos(2\phi')$ vanishes. Therefore, we are left with

$$Q(\hat{\boldsymbol{e}}_z, \boldsymbol{k} \| \hat{\boldsymbol{e}}_x) = \frac{15A\Theta_2(k)}{2} \int_0^\pi d\theta' \sin^5\theta' \int_0^{2\pi} d\phi' \cos^2\phi' \cos 2\phi'. \tag{10.24}$$

The $\phi'$ integral is $\pi/2$, while the $\theta'$ integral is $16/15$. So

$$Q(\hat{\boldsymbol{e}}_z, \boldsymbol{k} \| \hat{\boldsymbol{e}}_x) = 4\pi A\Theta_2(k). \tag{10.25}$$

We have now made part of the connection between polarization—represented by $Q$ here—and the formalism of anisotropies—described by $\Theta_l$ in general and $\Theta_2$ specifically for the quadrupole. This expression though applies only in the very simple case that the wavevector points along the $x$-axis, perpendicular to the line of sight $\hat{\boldsymbol{e}}_z$.

Let us generalize this expression to wavevectors pointing in an arbitrary direction in the $x$–$z$ plane: $\hat{\boldsymbol{k}} = (\sin\theta_k, 0, \cos\theta_k)$. In this case, $\mathcal{P}_2(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}')$ involves

$$(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}')^2 = \sin^2\theta_k \sin^2\theta' \cos^2\phi' + 2\sin\theta_k \cos\theta_k \cos\theta' \sin\theta' \cos\phi' + \cos^2\theta_k \cos^2\theta'. \tag{10.26}$$

The first term is identical to the $\hat{\boldsymbol{k}} \| \hat{\boldsymbol{e}}_x$ case just derived, multiplied by $\sin^2\theta_k$. The second and third terms both vanish after performing the $\phi'$ integral against $\cos(2\phi')$. Therefore,

$$Q(\hat{\boldsymbol{e}}_z, \boldsymbol{k} \perp \hat{\boldsymbol{e}}_y) = 4\pi A \sin^2\theta_k \Theta_2(k). \tag{10.27}$$

In Exercise 10.3 you will show that there is no $U$-polarization if $\boldsymbol{k}$ is in the $x$–$z$ plane: the polarization is all $Q$.

Now we can make use of what we have learned about the $E/B$-decomposition of polarization in Eqs. (10.11)–(10.12): the sky plane is the $x$–$y$ plane (since it is orthogonal to the direction of observation, the $z$ axis), so that a $\boldsymbol{k}$ vector in the $x$–$z$ plane lies along the $x$-axis on the sky. Since we find the polarization to be pure $Q$ in this case, we thus conclude

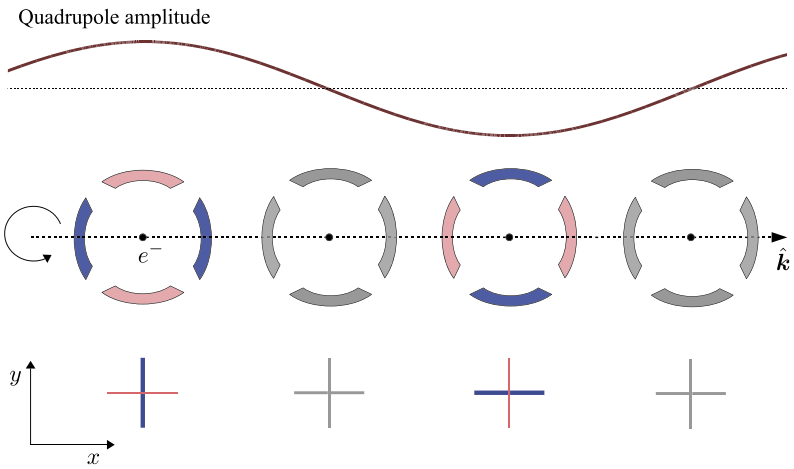



**FIGURE 10.8** Polarization pattern generated by a scalar perturbation (top panel) aligned with the $x$-axis. This scalar perturbation generates a spatially varying quadrupole in the temperature that is illustrated in the second panel through the temperature pattern seen by the electrons (indicated as dots). In three dimensions (with the $z$-axis coming out of the page) you should imagine this temperature pattern as azimuthally symmetric around $\hat{\boldsymbol{k}}$, i.e. the $x$-axis, as indicated by the circular arrow on the left. Now, suppose we observe the light scattered by the electrons in the $z$-direction coming out of the page. Following Sect. 10.2, the observed polarization pattern will appear as shown in the bottom panel. Comparing with Fig. 10.3, we can unambiguously tell that this is an $E$-mode.

that all polarization from scalar perturbations is in the $E$ component. In Exercise 10.2, you can derive the result for a general direction of $\boldsymbol{k}$, showing that it indeed obeys the expected $E$-mode form of Eq. (10.11):

$$\begin{aligned} Q(\hat{\boldsymbol{e}}_z, \boldsymbol{k}) &= 4\pi A \sin^2\theta_k \cos(2\phi_k)\Theta_2(k) \\ U(\hat{\boldsymbol{e}}_z, \boldsymbol{k}) &= 4\pi A \sin^2\theta_k \sin(2\phi_k)\Theta_2(k). \end{aligned} \tag{10.28}$$

Finally, we can generalize our result to an outgoing photon direction (line of sight) that is not necessarily along the $z$-axis, but given by a unit vector $\hat{\boldsymbol{n}}$. First, we use the fact that the polarization is completely described by the scalar $E$-mode on the sky, which is invariant under a rotation of $\hat{\boldsymbol{n}}$. We then only have to replace $\cos\theta_k$ with $\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{k}}$, so that

$$E(\hat{\boldsymbol{n}}, \boldsymbol{k}) = 4\pi A \left(1 - [\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{k}}]^2\right)\Theta_2(k). \tag{10.29}$$

At this point we should be mindful that our results throughout have been derived in the flat-sky approximation; the expressions on the full sky become quite a bit more complicated, without adding any new physics, however.

Fundamentally, the reason that scalar perturbations produce only $E$-modes is due to symmetry (see Fig. 10.8): if we consider a setup with a single plane-wave scalar perturbation along the $x$-axis, then we have azimuthal symmetry around this axis. The quadrupole then corresponds to the radiation being hotter (say) in the $\pm\hat{\boldsymbol{e}}_x$-directions, and colder

perpendicular to the $x$-axis (this is precisely described by the combination of $Y_{2,\pm 2}$ we encountered; see also Fig. C.1). As Fig. 10.8 shows, this azimuthal symmetry requires that the polarization pattern generated by Compton scattering is aligned with the $x$-axis as seen in Fig. 10.3, i.e. it is a pure $E$-mode. We will see below why this is so important. The overview article by Hu and White (1997) has many more detailed explanations and illustrations on the geometry of generating $E$- and $B$-mode polarization, going beyond the flat-sky approximation made here.

## 10.4 Boltzmann solution

In Sect. 5.7, we briefly introduced the polarization strength $\Theta_P(k, \mu, \eta)$. Now we understand how to define it precisely: it corresponds to the $Q$ polarization for a wavevector $\boldsymbol{k}$ along the $x$-direction. More generally, $\Theta_P = \Theta_E$ gives the amplitude of the $E$-mode polarization. Because of the azimuthal symmetry around the $\hat{\boldsymbol{k}}$-axis for scalar perturbations, the polarization is completely determined by a single function $\Theta_P(k, \mu, \eta)$. To make predictions for the observed $E$-mode, we need to derive and solve the Boltzmann equation that determines the evolution of $\Theta_P$. Fortunately, we have already done the hard work.

The left-hand side of the equation is determined by free-streaming and thus is the same as for $\Theta(k, \mu, \eta)$. The right-hand side contains the source and loss terms. We found in the previous section (Eq. (10.27)) that the outgoing polarization (for $\hat{\boldsymbol{k}}$ in the $x$–$z$ plane) was proportional to $(1-\mu^2)\Theta_2$ where $\mu$ is the cosine of the angle between $\hat{\boldsymbol{k}}$ and $\hat{\boldsymbol{n}}$. The proportionality constant has to contain the number of scattering events per $\eta$ interval, so $A \propto -\tau'$ (recall that the optical depth is measured going backward in time by convention, hence the minus sign). Thus, we expect a source term for $\Theta_P$ proportional to $-\tau'(1-\mu^2)\Theta_2$. Another important fact is that, if polarization is not sourced, the radiation becomes gradually unpolarized through Compton scattering. Hence, we expect a loss term proportional to $\Theta_P$, i.e. $\tau'\Theta_P$. Putting all of this together yields

$$\Theta_P' + ik\mu\Theta_P = -\tau'\left[b(1-\mu^2)\Theta_2 - \Theta_P\right], \tag{10.30}$$

where $b$ is a constant that we expect to be of order one. This equation gets almost everything right. One further effect to include is the polarization of the incoming radiation, which we have assumed to be unpolarized in our derivation. The result is (Bond and Efstathiou, 1987)

$$\Theta_P' + ik\mu\Theta_P = -\tau'\left[-\Theta_P + \frac{3}{4}(1-\mu^2)\Pi\right] \tag{10.31}$$

where

$$\Pi(k, \eta) \equiv \Theta_2 + \Theta_{P2} + \Theta_{P0}. \tag{10.32}$$

Now let us solve the Boltzmann equation for the polarization. In analogy to Eq. (9.49), the formal solution to Eq. (10.32) is

$$\Theta_P(k,\mu) = \int_0^{\eta_0} d\eta e^{ik\mu(\eta-\eta_0)-\tau(\eta)} S_P(k,\mu,\eta), \tag{10.33}$$

where the source term is

$$S_P(k,\mu,\eta) = -\frac{3}{4}\tau'(1-\mu^2)\Pi. \tag{10.34}$$

Using the definition in Eq. (9.56) of the visibility function, $g(\tau) = -\tau' e^{-\tau}$, this can be written as

$$\Theta_P(k,\mu) = \frac{3}{4}(1-\mu^2)\int_0^{\eta_0} d\eta\, g(\eta) e^{ik\mu(\eta-\eta_0)}\Pi(k,\eta). \tag{10.35}$$

A reasonable approximation is to assume that we can evaluate the integrand, except for the rapidly changing visibility function, at the time of decoupling. This leaves an integral over $g(\eta)$, which is unity. The result is

$$\Theta_P(k,\mu) \simeq \frac{3}{4}\Pi(k,\eta_*)(1-\mu^2)e^{ik\mu(\eta_*-\eta_0)}. \tag{10.36}$$

Neglecting $\eta_*$ compared with $\eta_0$ and rewriting the factors of $\mu$ as derivatives leads to

$$\Theta_P(k,\mu) \simeq \frac{3}{4}\Pi(k,\eta_*)\left(1+\frac{\partial^2}{\partial(k\eta_0)^2}\right)e^{-ik\eta_0\mu}. \tag{10.37}$$

To get the moments $\Theta_{P,l}$, we must multiply Eq. (10.37) by $\mathcal{P}_l(\mu)$ and integrate over all $\mu$ as in Eq. (5.66). This gives (Eq. (C.15))

$$E_l(k) = \Theta_{P,l}(k) \simeq \frac{3}{4}\Pi(k,\eta_*)\left(1+\frac{\partial^2}{\partial(k\eta_0)^2}\right) j_l(k\eta_0). \tag{10.38}$$

We identify $\Theta_{P,l}$ with $E_l$ because scalar perturbations generate only the $E$-mode.

Eq. (10.38) contains $(j_l + j_l'')(k\eta_0)$, which, using the spherical Bessel equation (Eq. (C.13)), can be rewritten as

$$j_l + j_l'' = -\frac{2}{k\eta_0} j_{l-1} + \frac{2(l+1)}{(k\eta_0)^2} j_l + \frac{l(l+1)}{(k\eta_0)^2} j_l. \tag{10.39}$$

Of the three terms on the right, the last one dominates on small scales. To see this, remember that the spherical Bessel function peaks roughly at $k\eta_0 \sim l$. For our order-of-magnitude estimate, this means that we can take $k\eta_0$ to be of order $l$ in the three terms on the right-hand side. The first and second then are of order $l^{-1}$, while the last is of order $l^2/(k\eta_0)^2 \sim 1$, so it dominates. Therefore,

$$E_l(k) \simeq \frac{3}{4}\Pi(k,\eta_*)\frac{l^2}{(k\eta_0)^2} j_l(k\eta_0). \tag{10.40}$$

In the tight-coupling limit, we can express $\Pi$ in terms of the quadrupole. As you can show in Exercise 10.5, $\Pi = 5\Theta_2/2$ for tight coupling. Therefore, the polarization moments today are

$$E_l(k) \simeq \frac{15}{8}\Theta_2(k,\eta_*)\frac{l^2}{(k\eta_0)^2}j_l(k\eta_0). \tag{10.41}$$

We can go one step further by noting that—in the tightly-coupled limit—the quadrupole is proportional to the dipole (Eq. (9.36)). Therefore,

$$E_l(k) \simeq -\frac{5k}{6\tau'(\eta_*)}\Theta_1(k,\eta_*)\frac{l^2}{(k\eta_0)^2}j_l(k\eta_0). \tag{10.42}$$

Eq. (10.42) is the expression for the polarization moments today induced by a single plane-wave scalar perturbation assuming the tightly-coupled limit. Three features are worthy of note. First, and most important, the polarization spectrum is seen to be smaller than the anisotropy spectrum by a factor of order $k/\tau'$ at the time of decoupling. This is a direct result of the twin facts that polarization is generated by the quadrupole moment and the quadrupole is suppressed in the early universe due to Compton scattering. Second, we expect there to be oscillations in the polarization power spectrum because $E_l \propto \Theta_1$, which undergoes acoustic oscillations. More quantitatively, we expect the polarization oscillations, just like the dipole, to be out of phase with the monopole $\Theta_0$. The peaks and troughs in the temperature anisotropy spectrum, arising primarily from oscillations in the monopole, should then be out of phase with the peaks and troughs in the polarization power spectrum. Finally, there is no analogue here to the integrated Sachs–Wolfe effect which impacts the temperature anisotropy spectrum. Polarization is not induced or modified by photons moving through changing (weak) gravitational potentials. Therefore, the polarization spectrum today is in some sense a more pristine view of the early universe, uncontaminated by later developments.

## 10.5 Polarization power spectra

Eq. (10.42) is an expression for the polarization moments from a single plane wave. The real universe contains not just one plane wave, but a superposition of many waves, all with differing amplitudes $\Theta_P(\boldsymbol{k},\hat{\boldsymbol{n}})$. The angular power spectrum of $E_l$ from a superposition of plane waves follows from the identical calculation for the temperature anisotropies; so, in analogy to Eq. (9.72), we define

$$\mathcal{T}_l^E(k) \equiv \frac{E_l(k)}{\mathcal{R}(\boldsymbol{k})}, \tag{10.43}$$

so that the polarization power spectrum becomes

$$C_{EE}(l) = \frac{2}{\pi}\int_0^\infty dk\, k^2 |\mathcal{T}_l^E(k)|^2 P_{\mathcal{R}}(k). \tag{10.44}$$

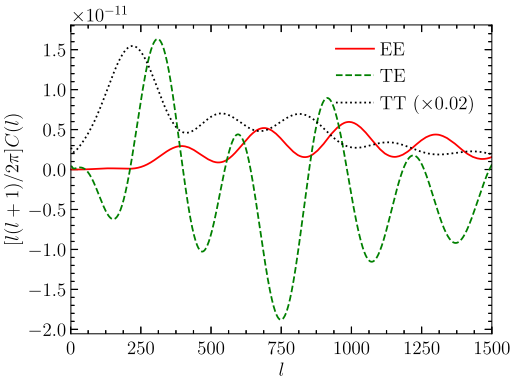



**FIGURE 10.9** Angular power spectra of temperature and $E$-mode polarization, as well as their cross-correlation, for the fiducial concordance cosmology. The polarization power spectra are smaller than those of the temperature by roughly a factor $1/50$, but less damped toward smaller scales.

We also know that

$$C_{BB}(l) = 0 \tag{10.45}$$

for scalar perturbations. Fig. 10.9 shows the polarization power spectrum $C_{EE}(l)$ from the full numerical calculation. Also shown is $C_{TT}(l)$ for comparison (dotted line; scaled down by a factor of 50).

As we expected from the tight-coupling approximation, the acoustic oscillations are both more pronounced in $C_{EE}(l)$, and out of phase with $C_{TT}(l)$; both facts are explained by $C_{EE}(l)$ being determined by the dipole of the photon distribution (recall that the quadrupole is related to the dipole in the tight-coupling approximation), while $C_{TT}(l)$ is a combination of monopole and dipole, where the monopole makes up the dominant contribution. Moreover, the dipole is less affected by photon diffusion, so that the polarization power spectrum is less damped on small scales, which can also be gleaned from the figure.

Fig. 10.9 also shows the cross-correlation between temperature and polarization, which is obtained by combining the temperature and $E$-mode polarization results:

$$C_{TE}(l) = \frac{2}{\pi}\int_0^\infty dk\, k^2 \left|\mathcal{T}_l^*(k)\mathcal{T}_l^E(k)\right| P_{\mathcal{R}}(k). \tag{10.46}$$

Note that Eq. (10.44) and Eq. (10.46) remain valid beyond the flat-sky approximation, although the calculation for $\Theta_l^E(k)$ becomes somewhat more involved.

The anticorrelation of $T$ and $E$ seen for $l \lesssim 200$ (i.e. $C_{TE}(l) < 0$) is of significance: it is a direct consequence of the fact that the initial conditions were setup outside the horizon so that only the cosine mode of the acoustic oscillations was generated (Eq. (9.26) and Eq. (9.28)). If both sine and cosine modes were generated, then there would not be a series of coherent peaks and troughs in the either the temperature or polarization spectrum (Dodelson, 2003). The acoustic oscillations are of course well-observed in the temperature spectrum. However, the first acoustic peak is at $\ell \sim 200$, corresponding to physical

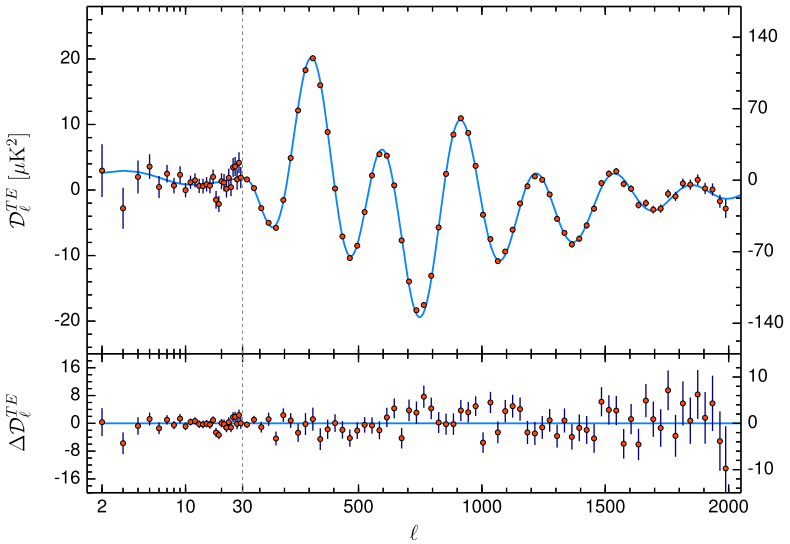



**FIGURE 10.10** *Upper panel:* Angular cross power spectrum of temperature and $E$-mode polarization as measured by the Planck experiment (points; $\mathcal{D}_l^{TE} \equiv l(l+1)C_{TE}(l)T_0^2/2\pi$ in analogy to Fig. 1.10); note that the scales of both $x$ and $y$ axes change at $l = 30$. The solid line shows the prediction for the fiducial concordance cosmology. Notice the negative $C_{TE}(l)$ in the range $30 \lesssim l \lesssim 200$. *Lower panel:* difference between theory and measurements. From Planck Collaboration (2018b).

scales that were within the horizon at recombination. One might come up with a viable model that generated perturbations around the time of recombination, taking care to ensure that the model generated only the cosine mode. The dip in the $TE$ spectrum though begins at $\ell \sim 30$, corresponding to scales that were still outside the horizon at the time of recombination. Therefore, any model that generates perturbations inside the horizon simply would not be able to explain the nonzero $C_{TE}(\ell)$ on those scales. Observation of this feature constitutes strong evidence for a much earlier origin of the initial perturbations, and inflation provides a natural explanation for this. Fig. 10.10 shows that this signature is by now impressively well measured in the observed CMB.

Both $C_{EE}(l)$ and $C_{TE}(l)$ contain valuable information about cosmological parameters. In particular, the $\mathcal{A}_s$–$\tau_{\rm rei}$ degeneracy described in Sect. 9.7.2 due to the late-time scattering of CMB photons after reionization is broken by polarization measurements. Recall our discussion in Sect. 9.7.2: on scales smaller than the horizon at the time of reionization (when electrons began to scatter again), which corresponds to $l \gtrsim 100$, the temperature anisotropies $\Theta$ are damped by a factor $e^{-\tau_{\rm rei}}$, while they are unaffected on the very largest scales. In polarization, this late epoch of scattering in fact *generates* anisotropies. Again, this works by converting the temperature quadrupole seen by the free electrons after reionization into polarized emission via Compton scattering. Eq. (10.41) still applies as long as we replace $\eta_*$ with $\eta_{\rm rei}$, and $\eta_0$ with $\eta_0 - \eta_{\rm rei}$, where $\eta_{\rm rei}$ is the conformal time when reionization happens. Because the photons free-stream between recombination and reionization,

the quadrupole $\Theta_2(k, \eta_{\rm rei})$ at reionization is suppressed if $k$ is inside the horizon at that time, i.e. if $k \gg aH(\eta_{\rm rei})$. For this reason, the polarization signal from reionization is only significant on very large angular scales; it is visible as the bump at $l < 10$ in Fig. 10.10, which is even more prominent in $C_{EE}(l)$. The amplitude of the reionization bump in $C_{EE}(l)$ is proportional to $\tau_{\rm rei}^2$, allowing for $\tau_{\rm rei}$ to be constrained independently of the temperature $C(l)$.

## 10.6 Detecting gravitational waves

There is a fundamental difference between the scalar perturbations we have considered in the previous sections and tensor perturbations. A scalar plane-wave perturbation has one direction associated with it: the direction of the wavevector $\boldsymbol{k}$. Once this direction is specified, all photon moments depend only on the angle between the photon momentum and the wavevector. If this angle is specified, there is an azimuthal symmetry about the $\hat{\boldsymbol{k}}$-direction. As we have seen in Fig. 10.8, this symmetry is the reason that only the $E$-mode is produced by scalar perturbations. There are two directions in a polarization field: $(i)$ the orientation of the polarization and $(ii)$ the direction in which the polarization strength is changing. For scalar perturbations, we saw in Fig. 10.3 that these directions must be aligned (or perpendicular to each other). Intuitively, each direction looks to the only vector it knows—$\boldsymbol{k}$—for guidance. This alignment is the salient characteristic of the $E$-mode.

The photon distribution induced by tensor perturbations is not rotationally symmetric about the $\boldsymbol{k}$-direction, a consequence of the characteristic pattern of the tensor metric perturbation shown in Fig. 6.1. Instead they induce an azimuthal dependence to the photon distribution. Recall from Eq. (6.85) derived in Exercise 6.14 that the resultant distribution varies as $\sin(2\phi)$ or $\cos(2\phi)$, where $\phi$ is the azimuthal angle about the $\boldsymbol{k}$-axis. This dependence on $\phi$ means that there is an additional direction to choose from when the polarization field is generated. We might expect then that the orientation of the polarization will not necessarily be aligned with the direction of changing polarization strength. That is, we might expect that gravitational waves will produce $B$-mode polarization, in addition to $E$-modes. This is exactly what we will show in this section.

Before working through the algebra, we should pause to understand the importance of the $B$-mode generated by tensor perturbations. Let us start with the difficulty of detecting tensors through either the temperature anisotropies or the $E$-mode. Both scalars and tensors contribute to the temperature and $E$-mode, so the only way to disentangle them is to take advantage of differences in their spectra as a function of $l$. We saw in the case of temperature anisotropies that this is a tricky game, though, because there are a number of free cosmological parameters that can be tweaked. So even if we had perfect knowledge of $C_{EE}(l)$ (without any noise; though this is clearly impossible due to cosmic variance), we would still not be able to tell unambiguously whether tensors were present. The $B$-mode is different. There is no contamination from scalar perturbations, so if we observe a $B$-mode in polarization, we know that it comes from gravitational waves. In

principle, this realization has unlimited power: no matter how small the tensor signal from inflation (no matter how small $H_{\rm inf}/m_{\rm Pl}$), we can ultimately detect this signal by searching for a $B$-mode, since there is no other component whose fluctuations could limit us. In practice, there are contaminants due to the polarization of foreground emission (due to dust and synchrotron radiation from the Milky Way) as well as nonlinear effects, in particular gravitational lensing (Sect. 13.3). Both of these contribute $B$-mode polarization. Nonetheless, CMB $B$-modes will be the most sensitive probe of primordial tensor modes for the foreseeable future.

The problem of computing the polarization pattern from a single plane-wave tensor perturbation is identical to that treated in Sect. 10.3. To find the outgoing polarization near the $z$-axis, we need to integrate over the incoming photon distribution. Really, our goal is to show that tensor modes produce $B$-mode polarization. This allows us to use a trick: Eq. (10.12) shows that, for a wavevector whose sky projection is along the $x$-axis, only the $B$-mode generates the $U$ polarization component. So, it suffices to show that tensor modes in this configuration produce $U$ via Eq. (10.20).

So let us choose the $\boldsymbol{k}$ vector as

$$\hat{\boldsymbol{k}} = \cos\alpha\,\hat{\boldsymbol{e}}_z + \sin\alpha\,\hat{\boldsymbol{e}}_x, \tag{10.47}$$

where $\alpha$ is the angle of $\hat{\boldsymbol{k}}$ with the line of sight $\hat{\boldsymbol{n}}$. For Eq. (10.20) we need to find the angular dependence of $\Theta^{\rm T}(\hat{\boldsymbol{n}}')$. This is the topic of Exercise 6.14 (in the case you have not gone through this exercise, we highly recommend that you do). There, you show that, for $\boldsymbol{k}$ lying along the $z$-axis, the angular dependence of the temperature anisotropy induced by a tensor perturbation is $\sin^2\theta'\cos(2\phi')$ (for $h_+$) or $\sin^2\theta'\sin(2\phi')$ (for $h_\times$). To be concrete, let us focus on $h_\times$; you can perform the same derivation for $h_+$ in Exercise 10.6. First, we can reexpress this angular dependence in terms of the unit vector $\hat{\boldsymbol{n}}'$ describing the direction of the incident photon:

$$\begin{aligned}\Theta^{\rm T}(\hat{\boldsymbol{n}}') \propto \sin^2\theta'\sin(2\phi') &= 2\sin^2\theta'\sin\phi'\cos\phi' \\ &= 2\hat{n}'_x\hat{n}'_y,\end{aligned} \tag{10.48}$$

where $\hat{n}'_{x,y,z}$ denote the components of the unit vector. Now we need to generalize this to a wavevector given by Eq. (10.47). For this, we need to rotate the coordinate system around the $y$-axis by $-\alpha$. This leaves $\hat{n}'_y$ unchanged, while $\hat{n}'_x$ changes according to

$$\hat{n}'_x \to \cos\alpha\,\hat{n}'_x - \sin\alpha\,\hat{n}'_z. \tag{10.49}$$

Now we are ready to plug the angular dependence into Eq. (10.20). The second term in Eq. (10.49) vanishes under the angular integrals, so we write only the first term, which, in terms of angles $\theta'$, $\phi'$, is in fact the same as before the rotation: $\propto \hat{n}'_x\hat{n}'_y = \sin^2\theta'\sin(2\phi')$. This

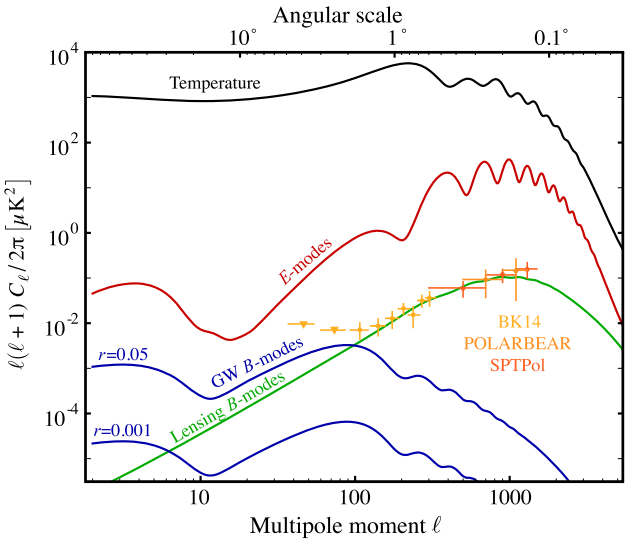



**FIGURE 10.11** $B$-mode CMB polarization anisotropy spectrum generated by tensor modes (or gravitational waves, GW), for two values of the tensor-to-scalar ratio $r = 0.001$, and $r = 0.05$, as labeled. For currently allowed values of $r$, the $B$-mode spectrum is even much smaller than the $E$-mode polarization. On small scales, $B$-mode polarization is generated by the gravitational lensing effect due to intervening structure acting on the primary CMB $E$-mode polarization. The experimental measurements, shown as points with error bars, have clearly detected this expected lensing signal. There also is an excess above this signal on large scales, which has been shown to be due to polarized foreground dust emission from the Milky Way. From Abazajian et al. (2016).

yields

$$U(\hat{\boldsymbol{e}}_z) \propto h_\times \cos\alpha \int_{-1}^{1} d\cos\theta' \, \sin^4\theta' \int_0^{2\pi} d\phi' \sin^2(2\phi')$$
$$= h_\times \cos\alpha \left(\frac{16}{15}\right)\left(\frac{\pi}{2}\right). \tag{10.50}$$

There we have it: a tensor mode which, projected on the sky, lies along the $x$-axis, produces $U$ polarization, which means it sources $B$-modes. To understand this a bit more intuitively, we go back to Fig. 10.8, whose middle panel shows the quadrupole of the temperature for a scalar perturbation. A tensor mode generates a quadrupole pattern that is rotated by 45° around the $z$-axis coming out of the page (see the pattern shown for $h_\times$ in Fig. 6.1, which is of essentially the form given in Eq. (10.48)); notice that this pattern only appears if $\hat{\boldsymbol{k}}$ has a nonzero $z$-component ($\alpha \neq \pi/2$), i.e. if it is not perfectly in the plane of the sky, which explains the $\cos\alpha$ factor in Eq. (10.50). This temperature quadrupole pattern means the polarization pattern shown in the bottom panel of Fig. 10.8 is likewise rotated by 45°. The comparison with Fig. 10.3 makes it plain that this is a $B$-mode.

Fig. 10.11 shows the $B$-mode power spectrum generated by inflationary gravitational waves, for two different values of the scalar-to-tensor ratio $r$ (Eq. (7.103); in fact, at fixed $\mathcal{A}_s$,

$C_{BB}(l)$ is linearly proportional to $r$). The signal that peaks around $l = 100$ comes from the recombination epoch which we have studied so far. In addition, there is another contribution on very large scales, peaking at $l < 10$, which is induced by the scattering off electrons after reionization. As we discussed, a similar feature, but due to scalar perturbations, appears in the $E$-mode power spectrum, allowing us to constrain $\tau_{\rm rei}$. Current upper limits, roughly of order $r < 0.05$, already constrain $C_{BB}(l)$ to be significantly smaller even than the E-mode polarization spectrum. Since there is no cosmic variance if there is no signal, the upper limits are controlled by the experimental sensitivity as well as removal of the foregrounds, so that future, more ambitious experiments can push down this limit further.

This is not quite the end of the story, however. As we will learn in Sect. 13.3, gravitational lensing acts on the CMB as well, by shuffling the observed locations of CMB hot and cold spots to slightly different positions. Shuffling the positions of a pure $E$-mode polarization pattern also generates $B$-modes. These "lensing $B$-modes" provide a guaranteed $B$-mode polarization power spectrum which has already been detected. Fortunately, as shown in Fig. 10.11, the angular power spectrum of these $B$-modes not only is of small amplitude compared to the linear scalar-induced CMB perturbations, but it also has a somewhat different shape from that of the primordial, tensor-mode-generated $B$-modes. We can then still search for the latter at least on large angular scales.

## 10.7 Summary

First, it is time to congratulate ourselves: we have successfully conquered the most technical derivation of this book! Fortunately, CMB polarization can be tackled largely with the same tools we have used throughout the book so far, namely the Boltzmann equation, aided by the simplification afforded by working on small angular scales (flat-sky approximation).

The power of CMB polarization stems from the fact that, being described by a symmetric, trace-free $2 \times 2$ matrix on the sky, it can be decomposed into two independent degrees of freedom: an $E$- (or gradient-) mode and a $B$- (curl-) mode, named after the analogy of their pattern with electrostatic and magnetic fields (see Fig. 10.4). Linear scalar perturbations produce only $E$-modes. We have shown this to be the case for CMB polarization, which is induced by Compton scattering of a local quadrupole temperature field. We will find the same to be true for galaxy shape correlations in Ch. 13, which will make abundant use of the results of Sect. 10.1.

The $E$-mode polarization of the CMB contains valuable cosmological information, and has been measured very precisely by now (Fig. 10.10). The negative cross-correlation with the temperature on large scales is another robust prediction of the inflationary scenario that is now confirmed. Further, the $E$-mode polarization at the lowest $l$ directly probes the late-time scattering of CMB photons after reionization, and can thus break the degeneracy between the amplitude of scalar perturbations and the optical depth $\tau_{\rm rei}$ due to late-time scattering that is present in the temperature alone.

$B$-mode polarization on the other hand can be used to search for non-scalar perturbations. As we have learned in Ch. 7, there is strong motivation to look for such perturbations, since inflation produces a background of gravitational waves. We showed in Sect. 10.6 that these tensor modes indeed produce $B$-mode polarization in the CMB. $B$-modes in fact are the currently most promising probe of this smoking-gun signature of inflation, with a corresponding level of experimental effort devoted to searching for them.

## Exercises

**10.1** Use Eq. (10.2) to derive how $I$, $Q$ and $U$ transform under a rotation of the coordinate system by an angle $\alpha$ around the line of sight $\hat{\boldsymbol{e}}_z$. Now do the same for $E$ and $B$, assuming a single plane wave with wavevector $\boldsymbol{l}$. Note that you will have to transform $\boldsymbol{l}$ as well. Finally, consider the transformation of all quantities under parity. In the 2D case, you can think of this as flipping *only* the $x$-axis.

**10.2** As the wavevector $\boldsymbol{k}$ moves out of the $x$–$z$ plane, show that the $Q$-polarization (for outgoing radiation in the $z$-direction) changes as $\cos(2\phi_k)$. To do this, first compute $\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}'$, and then integrate $\mathcal{P}_2(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}')$ over solid angle, with the weighting factor $\sin^2\theta' \cos(2\phi')$ derived in Eq. (10.19).

**10.3** This exercise focuses on the $U$-component of polarization from scalar perturbations.

**(a)** We showed that the $Q$-component of polarization from unpolarized incident radiation is given by Eq. (10.19), which stems from Eq. (10.15). The $Q$-component thus depends on the difference between $|\hat{\boldsymbol{\epsilon}}_i \cdot \hat{\boldsymbol{e}}_x|^2$ and $|\hat{\boldsymbol{\epsilon}}_i \cdot \hat{\boldsymbol{e}}_y|^2$. For the $U$-component, $\hat{\boldsymbol{e}}_x$ and $\hat{\boldsymbol{e}}_y$ here must be replaced by unit vectors rotated 45°, i.e., $(\hat{\boldsymbol{e}}_x + \hat{\boldsymbol{e}}_y)/\sqrt{2}$ and $(\hat{\boldsymbol{e}}_x - \hat{\boldsymbol{e}}_y)/\sqrt{2}$. With this replacement, derive Eq. (10.20).

**(b)** Show that a plane-wave perturbation with wavevector $\boldsymbol{k}$ lying in the $x$–$z$ plane does not produce any $U$-polarization in the outgoing $z$-direction.

**(c)** For the most general orientation of the wavevector,

$$\hat{\boldsymbol{k}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta), \tag{10.51}$$

show that $U$-polarization is given by Eq. (10.28).

**10.4** Draw the polarization patterns arising from a plane-wave scalar perturbation with (a) $\theta_k = \pi/8$, $\phi_k = \pi/8$; (b) $\theta_k = 3\pi/4$, $\phi_k = \pi/4$; (c) $\theta_k = 3\pi/4$, $\phi_k = 0$; and (d) $\theta_k = 3\pi/2$, $\phi_k = 0$. In each case, show that the direction of polarization is aligned with (or perpendicular to) the direction in which the polarization strength is changing.

**10.5** Find an expression for $\Pi \equiv \Theta_2 + \Theta_{P2} + \Theta_{P0}$ in the tight-coupling limit.

**(a)** When $\tau'$ is very large, the terms multiplying it on the right-hand side of Eq. (10.32) must cancel. Write down this equality for $\Theta_P(\mu)$ in terms of the moments, $\Theta_2$, $\Theta_{P2}$, and $\Theta_{P0}$.

**(b)** Expand $\Theta_P(\mu)$ in terms of Legendre polynomials, keeping only the monopole and the quadrupole. Then equate the coefficients of $\mathcal{P}_0$ and $\mathcal{P}_2$.

**(c)** This leads to two equations for three unknowns. Show that solving for the two polarization moments in terms of the temperature quadrupole gives $\Theta_{P0} = 5\Theta_2/4$ and $\Theta_{P2} = \Theta_2/4$.

**(d)** Use the results of (c) to determine $\Pi$ in terms of $\Theta_2$.

**10.6** In the text we proved that tensor modes generate $B$-mode polarization by studying the tensor-mode component $h_\times$. Repeat the derivation for the other component $h_+$, starting from Eq. (10.47). Can you make sense of your results? (hint: study the discussion after Eq. (10.50) and compare the two component patterns in Fig. 6.1.)

# 11

# Probes of structure: tracers

In the previous two chapters, we saw that a wealth of information can be extracted from the angular power spectrum $C(l)$ of the CMB temperature, as well as those of the polarization. On the other hand, Ch. 8 supplied us with a similarly precise prediction for the linear matter power spectrum $P_{\rm L}(k,z)$, and showed that there is rich information in $P_{\rm L}(k,z)$ as well: in particular on the Hubble constant, dark energy, and the mass of neutrinos. Unlike the case for the CMB, however, we do not have a direct way of measuring the matter power spectrum; after all, the bulk of matter is in form of dark matter, and even much of the baryonic matter is not readily observable (e.g., dilute hot gas). However, there are many observables that indirectly probe the matter distribution. In this chapter, we will cover perhaps the most important one, galaxy clustering, which uses galaxies (more generally, any astrophysical objects) as *tracers* of the large-scale matter distribution. We will also cover another tracer that uses the CMB as a backlight. In later chapters, we will learn about further important probes, including galaxy clusters and gravitational lensing.

The most direct measurement of the galaxy density field is supplied by galaxy redshift surveys, wherein the angular positions and the redshifts (which are a measure of radial distance) of galaxies are recorded. We thus have a 3D position for each galaxy, which allows us to measure their three-dimensional statistics such as the galaxy power spectrum $P_{\rm g,obs}(\boldsymbol{k})$. There are, however, a number of problems with the interpretation of the galaxy power spectrum as measured from redshift surveys. First, there is the problem of *bias*, the fact that galaxy clustering is different from that of matter. Second, the galaxy redshifts contain not only the cosmological redshift, which is simply a function of their distance, but also a Doppler shift due to the peculiar velocities of galaxies. Recall that a galaxy's redshift is determined solely by the Hubble expansion (and hence redshift is a perfect indicator of distance) only if the galaxy is stationary on the comoving grid. Most galaxies are not stationary so have nonnegligible *peculiar velocities*. These velocities contribute to a galaxy's observed redshift via the Doppler effect. Thus, even an accurate measurement of a galaxy's redshift does *not* translate into an unambiguous measurement of its radial distance away from us. Moreover, galaxy velocities are not random, but correlate with the matter density field itself. This leads to modifications of galaxy statistics known as *redshift-space distortions (RSDs)*.

However, at first order, neither bias nor RSD change the shape of the large-scale clustering of galaxies, a fact that follows from simplifying assumptions on bias and galaxy velocities. In this chapter, we will just assert these assumptions; we will understand why they hold in Ch. 12. This fact allows us to use the shape of the galaxy power spectrum to constrain the expansion history, by using the *baryon acoustic oscillation (BAO)* feature in the matter power spectrum. This makes BAO one of the major probes of dark energy.

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00017-6
Copyright © 2021 Elsevier Inc. All rights reserved.

The apparent nuisance of RSD also has a virtue: by measuring galaxy velocities through their Doppler-induced statistical effect on the galaxy power spectrum, we can measure the rate at which structure grows. This is another probe of dark energy, and in fact gravity.

While galaxy redshift surveys are rich in information, they are also expensive, due to the simple fact that obtaining redshifts is time consuming: it is much easier to get the angular positions of galaxies than it is to also measure redshifts. To obtain the redshift, one has to take a spectrum of each galaxy, which requires a significantly larger number of photons than are necessary for imaging the galaxy. Photometric surveys, which merely take images of parts of the sky, compensate for the lack of radial information by observing many more galaxies. Moreover, surveys focused on measuring gravitational lensing also yield the angular positions of galaxies as a byproduct. Clearly, then, one skill we must acquire is the ability to make predictions about the angular galaxy power spectrum $C_{\rm g}(l)$. In Sect. 11.2 we will see that $C_{\rm g}(l)$ is an integral over the 3D galaxy power spectrum.

A similar integral, but over the 3D power spectrum of pressure perturbations in the ionized gas, is the expression we will derive for a different observable in Sect. 11.3. In the late universe, the temperature of the ionized gas is far hotter than that of the CMB. Scattering of the CMB off these hot electrons tends to shift the CMB photons to higher energies, leading to a characteristic distortion of the CMB black-body spectrum. This so-called *Sunyaev–Zel'dovich (SZ)* signal can be isolated from the primordial CMB, and hence be used for an integral measurement of pressure in the ionized universe, as well as to search for massive galaxy clusters.

## 11.1 Galaxy clustering

Fig. 11.1 shows the positions of galaxies in a slice through the volume mapped by the Sloan Digital Sky Survey (SDSS). In a redshift survey such as this, which contains close to a million galaxies in total, what statistic can we compute that can be compared with theory? Similar to the case of the CMB, the simplest statistic is the now three-dimensional power spectrum $P_{\rm g,obs}(\boldsymbol{k})$.

First let us introduce some notation. Fig. 11.2 shows the geometry: a given galaxy is at a comoving distance $\chi(z)$ (Eq. (2.34)) from us. Therefore, the three-dimensional position vector $\boldsymbol{x}$ has components

$$\boldsymbol{x}_{\rm obs}(z,\theta,\phi) = \chi(z)\,\hat{\boldsymbol{n}}(\theta,\phi); \quad \hat{\boldsymbol{n}} = \frac{\boldsymbol{x}_{\rm obs}}{|\boldsymbol{x}_{\rm obs}|}. \tag{11.1}$$

The unit vector $\hat{\boldsymbol{n}}$ is in one-to-one correspondence with the galaxy's position on the sky, which in turn is specified by the two angles $\theta, \phi$. Moreover, in an unperturbed universe, the distance $\chi$ of the galaxy is directly related to the observed redshift $z$ by $\chi(z)$ (see Sect. 2.2). In order to compute this function, we need to assume an expansion history for the universe, which is part of what we want to measure. In an actual survey analysis, one typically assumes a fiducial cosmology with distance-redshift relation $\chi_{\rm fid}(z)$, which in general dif-

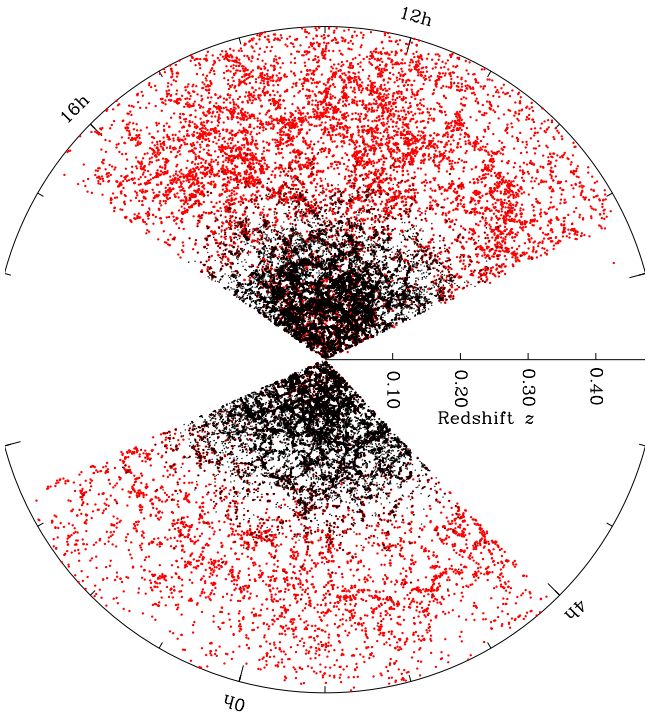



**FIGURE 11.1** The distribution of galaxies measured in the SDSS survey (more precisely, those within a slice of $\pm 3$ deg of the celestial equator). The colors denote different galaxy samples: magnitude-limited main sample (black, as shown in Fig. 1.8) and luminous red galaxy sample (LRG, red (light gray in print version)). Notice that the LRG sample covers a much greater volume, since it consists of luminous galaxies observable to larger distances. Image Credit: Michael Blanton and the SDSS Collaboration.

fers from the one of the real universe:

$$\chi_{\rm fid}(z) = \chi(z) + \delta\chi(z). \tag{11.2}$$

A second problem with Eq. (11.1) is that $|\boldsymbol{x}_{\rm obs}|$ no longer corresponds to the true distance of the galaxy if the galaxy is not at rest with respect to the background universe. This is simply due to the fact that the observed redshift of a galaxy is given by

$$1 + z = \frac{1}{a_{\rm em}}\left[1 + u_\parallel\right], \quad u_\parallel = \boldsymbol{u}_{\rm g}\cdot\hat{\boldsymbol{n}}, \tag{11.3}$$

where $a_{\rm em}$ is the scale factor at which the light from the galaxy was emitted. That is, the factor $1/a_{\rm em}$ is the cosmological redshift. The second term is the Doppler shift due to the peculiar velocity $\boldsymbol{u}_{\rm g}$ of the galaxy, at linear order in $u_{\rm g}$. Notice that $1/a_{\rm em}$ multiplies both the cosmological and Doppler term: that is, the *fractional* effect of a given peculiar velocity on the redshift is independent of how far away the galaxy is. Here, we have assumed that the velocities of galaxies are much smaller than the speed of light (see the discussion

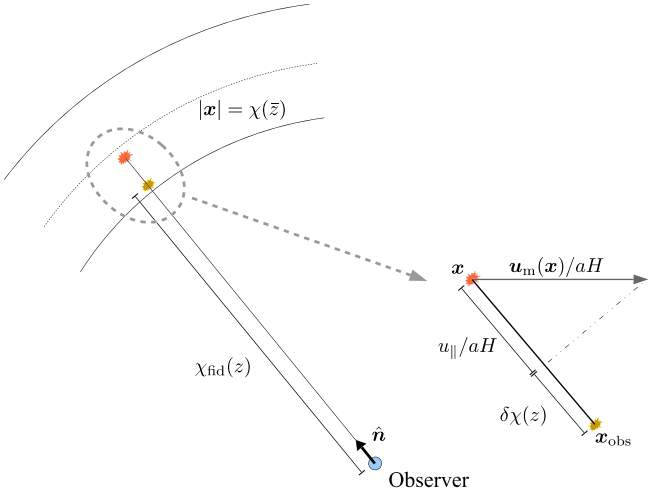



**FIGURE 11.2** Sketch illustrating the complications in measuring the 3D galaxy density field. Galaxies are selected within a narrow redshift slice (solid arcs) centered on $\bar{z}$ (dotted arc). A galaxy is observed at a position $\hat{\boldsymbol{n}}$ in the sky, with redshift $z$. We assign it the apparent position $\boldsymbol{x}_{\rm obs}$, which differs from the true position $\boldsymbol{x}$ due to two reasons (see Eq. (11.5)): the distance-redshift relation assumed is not the true one, and the redshift includes a Doppler contribution due to the galaxy velocity along the line of sight, $u_\parallel$.

in Sect. 12.1), allowing us to stop at linear order in $u_{\rm g}$. In the following, we set the galaxy velocity to be equal to that of matter, $\boldsymbol{u}_{\rm g} = \boldsymbol{u}_{\rm m}$. We will see in Sect. 12.6 that this is indeed the case *on large scales*. Eq. (11.3) further neglects the Sachs–Wolfe and integrated-Sachs–Wolfe contributions to the redshift which are much smaller, and become important only on extremely large scales.

If $\boldsymbol{u}_{\rm m} = 0$, and the fiducial cosmology happens to be the true one, then our distance estimate $|\boldsymbol{x}| = \chi(z)$ is accurate. We can compute the error made in Eq. (11.1):

$$\Delta\boldsymbol{x}_{\rm RSD} = \left.\frac{\partial \boldsymbol{x}_{\rm obs}}{\partial u_\parallel}\right|_{u_\parallel=0} u_\parallel = \frac{1}{aH} u_\parallel \hat{\boldsymbol{n}} \tag{11.4}$$

where the subscript $_{\rm RSD}$ identifies this effect as due to redshift-space distortion. Because each galaxy is shifted along its line of sight, $\Delta\boldsymbol{x}_{\rm obs}$ is proportional to $\hat{\boldsymbol{n}}$. In other words, the observed (angular) position on the sky is unchanged. The sign in Eq. (11.4) is also clear: $u_\parallel > 0$ means that the galaxy is moving away from us, leading to an additional Doppler redshift, and hence an increase in the distance. Combining the effect of the wrong assumed cosmology, Eq. (11.2), with Eq. (11.4), and working to linear order in both $\delta\chi$ and $u_\parallel$, we obtain

$$\boldsymbol{x}_{\rm obs} = \boldsymbol{x} + \left(\delta\chi(z) + \frac{1}{aH} u_\parallel(\boldsymbol{x})\right)\hat{\boldsymbol{n}}, \tag{11.5}$$

where $\boldsymbol{x}$ is the true three-dimensional position of the galaxy (i.e. the one we would observe in the absence of velocities and given knowledge of the true expansion history). As we will

see, this shift in the observed coordinate from the true position of the galaxy has important implications for observations of galaxy clustering. While it will make our life in interpreting galaxy statistics a little more difficult, there is rich cosmological information in both sources of the coordinate shift. So it will be worth it.

### 11.1.1 Galaxy statistics

Suppose we measure the power spectrum in terms of the observed coordinates $\boldsymbol{x}_{\rm obs}$ (Eq. (11.5)). How is this distorted power spectrum related to the underlying "true" spectrum of the galaxies? The first studies of this problem go back to the 1970s. Here, we will follow the derivation of the most well-known paper on the topic (Kaiser, 1987), working within the context of linear theory. While Kaiser considered only RSD, we will also include the position shift due to the incorrect cosmology, since it allows for a unified derivation of all effects entering the observed galaxy power spectrum.

The starting point is the realization that the number of galaxies in a particular region is the same, whether we use the observed position $\boldsymbol{x}_{\rm obs}$ or the true coordinate $\boldsymbol{x}$. Imagine dividing the survey volume into many volume elements (voxels). We can count the galaxies in each volume element and use that to construct the galaxy density field $n_{\rm g,obs}(\boldsymbol{x}_{\rm obs})$; we could do the same in terms of the true galaxy positions, if we had access to them, yielding $n_{\rm g}(\boldsymbol{x})$. Since the number of galaxies is the same in each case, we have

$$n_{\rm g,obs}(\boldsymbol{x}_{\rm obs})d^3x_{\rm obs} = n_{\rm g}(\boldsymbol{x})d^3x \tag{11.6}$$

where $n_{\rm g}$ is the density of galaxies at $\boldsymbol{x}$ in real space, and $n_{\rm g,obs}$ is the density in redshift space. The infinitesimal volume element around a point in observed coordinates is

$$d^3x_{\rm obs} = x_{\rm obs}^2 dx_{\rm obs} d\Omega, \tag{11.7}$$

where $x_{\rm obs} \equiv |\boldsymbol{x}_{\rm obs}|$, while the volume around a point in real space is

$$d^3x = x^2 dx d\Omega. \tag{11.8}$$

The angular volume elements $d\Omega$ are identical, so

$$n_{\rm g,obs}(\boldsymbol{x}_{\rm obs}) = n_{\rm g}(\boldsymbol{x})J \tag{11.9}$$

where the Jacobian $J$ is given by

$$J \equiv \left|\frac{d^3x}{d^3x_{\rm obs}}\right| = \left|\frac{dx}{dx_{\rm obs}}\right| \frac{x^2}{x_{\rm obs}^2}. \tag{11.10}$$

With this and Eq. (11.5), the Jacobian is calculated at linear order in both $\delta\chi$ and $u_\parallel$ to be

$$J = \left(1 + \frac{\delta\chi}{x} + \frac{u_\parallel}{aHx}\right)^{-2} \left|1 + \frac{d}{dx}\delta\chi + \frac{1}{aH}\frac{\partial}{\partial x}u_\parallel\right|^{-1}. \tag{11.11}$$

Note the key difference between $\delta\chi$ and $u_\parallel$: while the former depends only on redshift (for which we can use $x$ as a proxy), $u_\parallel$ fluctuates as a function of the three-dimensional position. To compute the derivative of $\delta\chi$, we use

$$\frac{d}{dx}\delta\chi = \frac{dz}{dx}\frac{d\delta\chi}{dz} = H\delta(H^{-1}) = -H^{-1}\delta H, \tag{11.12}$$

where $\delta H(z) = H(z) - H_{\rm fid}(z)$ is the difference between the fiducial and true Hubble rates, and we have used $dz/dx = dz/d\chi = H$ multiple times. Therefore, Eq. (11.11) becomes

$$\begin{aligned} J &= \left(1+\frac{\delta\chi}{x}+\frac{u_\parallel}{aHx}\right)^{-2}\left|1-H^{-1}\delta H+\frac{1}{aH}\frac{\partial}{\partial x}u_\parallel\right|^{-1} \\ &\simeq \left(1-2\frac{\delta\chi}{x}+H^{-1}\delta H-2\frac{u_\parallel}{aHx}\right)\left(1-\frac{1}{aH}\frac{\partial}{\partial x}u_\parallel\right) \end{aligned} \tag{11.13}$$

where the second line discards some second-order terms and moves the $\delta H$ term into the first parentheses. To see why, consider this term and the $\delta\chi$ before it. These depend only on $|\boldsymbol{x}|$ via the redshift $z$. If we imagine looking at galaxies in a fairly narrow redshift slice around $\bar{z}$, as is usually done to avoid an evolving galaxy population, then we can set $x$ to $\bar{\chi} = \chi(\bar{z})$, the distance to the mean $\bar{z}$ of this redshift slice. Further, we can evaluate $\delta\chi$, $\delta H$, and $H$ at $\bar{z}$. With this, these terms simply become a constant prefactor.

Next, consider the $u_\parallel/aHx$ term. As Kaiser realized, this contribution is small in most practical cases. $u_\parallel/aH$ is the apparent displacement of galaxies due to their line-of-sight velocity; using linear theory and plugging in numbers, one finds that this displacement is typically $\lesssim 10\,h^{-1}$ Mpc (Exercise 11.3). On the other hand, $x \sim \bar{\chi}$ is at least many hundreds of Mpc in state-of-the-art galaxy surveys. Thus, this term is small and can be neglected. The same is *not* true of the term involving $\partial u_\parallel/\partial x$: since the velocity field fluctuates, this term is actually quite large. How large precisely, we will see in a minute. We finally arrive at our simplified Jacobian:

$$J \simeq \bar{J}\left(1-\frac{1}{aH}\frac{\partial}{\partial x}u_\parallel\right); \quad \bar{J} = 1-2\frac{\delta\chi(\bar{z})}{\bar{\chi}}+H^{-1}(\bar{z})\delta H(\bar{z}). \tag{11.14}$$

The number densities in true and observed coordinates are $n_{\rm g} = \bar{n}_{\rm g}(1+\delta_{\rm g})$ and $n_{\rm g,obs} = \bar{n}_{\rm g}(1+\delta_{\rm g,obs})$, respectively, with $\bar{n}_{\rm g}$ the average number density. In practice, the mean $\bar{n}_{\rm g}$ is determined by counting all galaxies in the redshift slice and dividing by its volume. This ensures that $\delta_{\rm g,obs}$, averaged over the survey, has vanishing mean as desired. In light of Eq. (11.9), and expanding to first order in perturbations, the observed galaxy overdensity is

$$1+\delta_{\rm g,obs}(\boldsymbol{x}_{\rm obs}) = \bar{J}\left[1+\delta_{\rm g}(\boldsymbol{x}[\boldsymbol{x}_{\rm obs}])-\frac{1}{aH}\frac{\partial}{\partial x}u_\parallel(\boldsymbol{x}[\boldsymbol{x}_{\rm obs}])\right]. \tag{11.15}$$

Note that the galaxy density and velocity on the right-hand side are evaluated at the true position. In the following two sections, we will deal separately with the effect of velocities, i.e. RSD, and the effect of the incorrect cosmology.

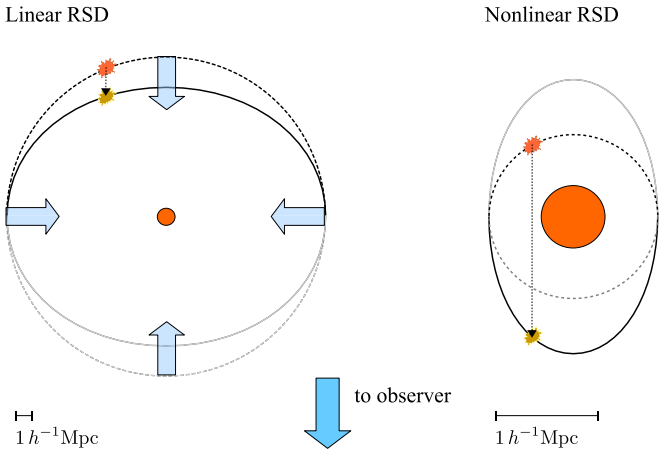



**FIGURE 11.3** Redshift-space distortions, in the linear/large-scale (left) and nonlinear/small-scale variants (right), both considering the case of a central overdensity denoted by the filled circle. The observer is assumed to be far away below the figure, so that the line-of-sight direction $\hat{\boldsymbol{n}}$ is vertical. In each case, a contour of constant density (dashed lines), which is circular in real space, is distorted in redshift space (solid lines) so that it looks asymmetric. Wide arrows indicate the direction of the velocity flow, while arrows with dashed lines indicate the displacement due to the line-of-sight velocity. In the nonlinear case, as the absolute scales are smaller, a point on the "far side" (top) of the overdensity is mapped onto a point on the opposite side.

### 11.1.2 Redshift-space distortions

Before we begin computing the galaxy power spectrum, let us think about what we qualitatively expect the effect of peculiar velocities on galaxy clustering to be. Fig. 11.3 illustrates the distortions that appear in redshift space. The left panel shows the large-scale case we are mostly interested in here. A large-scale overdense region, towards which surrounding galaxies are falling, appears squashed in redshift space: the galaxies closest to us are moving toward the center of the overdense region and hence away from us, so they appear farther from us (and closer to the center of the overdense region) than they actually are. Similarly, galaxies on the "other side" of the perturbation are moving toward us, so they appear closer to us than they actually are. The overall effect is to induce an apparent anisotropy in an otherwise circular overdensity. Since moving galaxies towards each other increases their number density (the effect captured by the Jacobian introduced in the previous section), we actually expect the clustering in redshift space to be *stronger* than in real space.

As we move to smaller, nonlinear scales, the nature of the redshift-space distortion changes. Velocities on small scales are typically a bit larger, but more importantly, the displacement into redshift space, $u_\parallel/aH$, becomes much larger compared to the distance separating the two galaxies which we are correlating, since clustering on small scales by definition means that we are considering pairs of galaxies that are closer together. The result is shown in the right panel of Fig. 11.3. The observed contour of constant density is very elongated along the line of sight. Moreover, the true and observed positions of the

galaxy have swapped places. This means that the quadrupole moment of the clustering has the opposite sign than it does in the linear case.

Notice that our assumption that we are looking at a small distortion of the true galaxy position cannot correctly describe the case shown in the right panel of Fig. 11.3. It is clear then that accounting for redshift-space distortions will be a tricky business, requiring careful treatment not only of linear overdensities, but also of the much more complicated effects of nonlinearities on smaller scales. We will content ourselves with a quantitative treatment of linear distortions in this chapter, since this applies on large scales and is the starting point for all further work.

Now let us move on to compute the power spectrum of galaxies from Eq. (11.15). Neglecting $\delta\chi$ and $\delta H$, we can set $\bar{J}$ to unity. We also need relations of $\delta_{\rm g}$ and $\boldsymbol{u}_{\rm m}$ to the matter density $\delta$. Let us begin with $\boldsymbol{u}_{\rm m}$, which lies along the $\hat{\boldsymbol{k}}$-direction. At late times ($z \lesssim 10$), baryons and CDM move together ($u_{\rm b} = u_{\rm c}$) and their overdensities are equal ($\delta_{\rm b} = \delta_{\rm c}$). We can then use the continuity equation (8.12) for both, i.e. for the total matter density:

$$\delta_{\rm m}{}' + i\boldsymbol{k}\cdot\boldsymbol{u}_{\rm m} = -3\Phi' \tag{11.16}$$

where primes denote derivatives with respect to $\eta$. Since we are working on sub-horizon scales, we can set the right-hand side to zero, as it is suppressed by $(aH/k)^2$ compared to the terms on the left-hand side. Using the fact that the time dependence of the linear density field is given by the growth factor $D_+(\eta)$, we can then solve for the velocity in terms of the density:

$$\boldsymbol{u}_{\rm m}(\boldsymbol{k},\eta) = \frac{i\boldsymbol{k}}{k^2}\frac{D_+{}'}{D_+}\delta(\boldsymbol{k},\eta) = aHf\frac{i\boldsymbol{k}}{k^2}\delta(\boldsymbol{k},\eta)\,, \tag{11.17}$$

where the linear growth rate $f = d\ln D_+/d\ln a$ is defined in Eq. (8.78). The growth rate is close to unity for a $\Lambda$CDM universe (and exactly 1 for a flat matter-dominated cosmology). There are two important points about the relation between the density and the velocity in Eq. (11.17) that should be emphasized. First, notice that the velocity in Fourier space is proportional to the wavevector $\boldsymbol{k}$; that is, it is a longitudinal vector. Physically, this corresponds to the absence of a curl component of the velocity. Indeed, we will see in Ch. 12 that the curl component decays rapidly (similar to the vector modes discussed in Ch. 6), so that this is an accurate assumption. The second point about the relation Eq. (11.17) between the velocity and the density is that it holds only in linear theory, and hence on very large scales only; we will introduce methods to go beyond this assumption in Ch. 12.

Next, in order to relate $\delta_{\rm g}$ to $\delta_{\rm m}$, we assume the *linear bias relation*

$$\delta_{\rm g}(\boldsymbol{x},\eta) = b_1(\eta)\delta_{\rm m}(\boldsymbol{x},\eta). \tag{11.18}$$

Because galaxies are complicated, highly nonlinear tracers of the large-scale structure, their density perturbation is not the same as that of matter. Why should it be simply linearly related though? The answer is that this is a guaranteed result *at linear order* in perturbations and on large scales. We will justify this in more rigor in Sect. 12.6. The bias

parameter $b_1$ depends sensitively on the galaxy sample considered and is in general redshift dependent. Finally, because galaxies are discrete tracers, the galaxy density field has noise. We will include this at the very end, since it is independent of RSD.

Making use of these results and Eq. (11.15) (with $\bar{J} = 1$), we see that the overdensity in redshift space is actually a sum of the overdensity in real space and a correction due to peculiar velocity,

$$\delta_{\rm g,RSD}(\boldsymbol{x}) = b_1\delta_{\rm m}(\boldsymbol{x}) - \frac{\partial}{\partial \boldsymbol{x}}\left[\frac{\boldsymbol{u}_{\rm m}(\boldsymbol{x})\cdot\hat{\boldsymbol{x}}}{aH}\right], \tag{11.19}$$

where we use the subscript $_{\rm RSD}$ since RSD is only one of two effects we need to include to get to $\delta_{\rm g,obs}$. Here and in the following, we suppress the time argument for clarity. We also replaced $\boldsymbol{x}_{\rm obs}$ with $\boldsymbol{x}$, since these positions differ by a term that is already a perturbation, and expanding the argument of $\delta_{\rm g}$ and $\boldsymbol{u}_{\rm m}$ would lead to higher-order terms, so we can neglect the distinction between $\boldsymbol{x}$ and $\boldsymbol{x}_{\rm obs}$. For clarity, we then simply use $\boldsymbol{x}$. Further, $aH$ is always evaluated at the mean redshift $\bar{z}$ and thus is effectively a constant (this is very accurate for a narrow redshift slice).

We now introduce the *distant-observer* approximation, which is essentially a flat-sky approximation. The idea is that we can treat the direction vector $\hat{\boldsymbol{n}} = \boldsymbol{x}/x$ as fixed, neglecting variations from galaxy to galaxy. This is justified if the angular scales involved are small (cf. Fig. 11.2). That is, as long as the galaxies are relatively close to each other in the $(x^1, x^2)$ plane, we can approximate $\hat{\boldsymbol{x}}\cdot\boldsymbol{u}_{\rm m} \to \hat{\boldsymbol{e}}_z\cdot\boldsymbol{u}_{\rm m}$, where $\hat{\boldsymbol{e}}_z$ is a radial vector pointing to the center of the sky area covered by the survey of interest (and we choose this to be the $z$-axis).

In the distant-observer approximation, we can directly compute the Fourier transform of the redshift-space overdensity

$$\begin{aligned}\delta_{\rm g,RSD}(\boldsymbol{k}) &= \int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\left[b_1\delta_{\rm m}(\boldsymbol{x}) - \frac{\partial}{\partial x}\left[\frac{\boldsymbol{u}_{\rm m}(\boldsymbol{x})\cdot\hat{\boldsymbol{e}}_z}{aH}\right]\right] \\ &= b_1\delta_{\rm m}(\boldsymbol{k}) - if\int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\frac{\partial}{\partial x}\left[\int\frac{d^3k'}{(2\pi)^3}e^{i\boldsymbol{k}'\cdot\boldsymbol{x}}\delta_{\rm m}(\boldsymbol{k}')\frac{\boldsymbol{k}'}{k'^2}\cdot\hat{\boldsymbol{e}}_z\right],\end{aligned} \tag{11.20}$$

the first equality following from our Fourier convention and Eq. (11.19) and the second from Eq. (11.17). The derivative with respect to $x$ acts on the exponential, bringing down a factor of $i\boldsymbol{k}'\cdot\hat{\boldsymbol{x}}$, which we again set to $i\boldsymbol{k}'\cdot\hat{\boldsymbol{e}}_z$, so

$$\delta_{\rm g,RSD}(\boldsymbol{k}) = b_1\delta_{\rm m}(\boldsymbol{k}) + f\int\frac{d^3k'}{(2\pi)^3}\delta_{\rm m}(\boldsymbol{k}')\left(\hat{\boldsymbol{k}}'\cdot\hat{\boldsymbol{e}}_z\right)^2\int d^3x e^{i(\boldsymbol{k}'-\boldsymbol{k})\cdot\boldsymbol{x}}. \tag{11.21}$$

The $\boldsymbol{x}$ integral yields $(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}'-\boldsymbol{k})$. Therefore, in the distant-observer approximation,

$$\delta_{\rm g,RSD}(\boldsymbol{k}) = \left[b_1 + f\mu_k^2\right]\delta_{\rm m}(\boldsymbol{k}). \tag{11.22}$$

Here $\mu_k$ is defined to be $\hat{\boldsymbol{e}}_z\cdot\hat{\boldsymbol{k}}$, the cosine of the angle between the line of sight and the wavevector $\hat{\boldsymbol{k}}$. Eq. (11.22) quantifies what we have anticipated about (large-scale) redshift-space distortions. First of all, since $f\mu_k^2 \geq 0$, the apparent overdensity in redshift space

is *larger* than in real space, where it would be simply $b_1\delta_{\rm m}(\boldsymbol{k})$. This can be gleaned from Fig. 11.3: if equal-density contours are squashed, then galaxies are moved closer together and hence their density is increased around an overdensity. The opposite happens around underdensities. Both effects act to increase the apparent galaxy density contrast. The second feature of Eq. (11.22) worth noting is that the enhancement is for perturbations with wavevector parallel to the line of sight. A perturbation with $\boldsymbol{k}$ perpendicular to the line of sight—one in which the density along the line of sight is constant—experiences no redshift-space distortion.

Correspondingly, the galaxy power spectrum in redshift space depends not only on the magnitude of $\boldsymbol{k}$ but also on its direction, which we are parameterizing with $\mu_k$. It follows from Eq. (11.22) that

$$P_{\rm g,RSD}(k,\mu_k,\bar{z}) = P_{\rm L}(k,\bar{z})\left[b_1 + f\mu_k^2\right]^2 + P_N, \tag{11.23}$$

where $P_{\rm L}(k,z)$ is the linear matter power spectrum studied in Ch. 8, and both $b_1$ and $f$ are understood to be evaluated at the mean redshift $\bar{z}$. We have finally allowed for a noise term in the galaxy power spectrum, which we assume to be "white," i.e. to be a scale-independent constant $P_N$. This holds, for example, if the galaxies are Poisson-sampled from an underlying continuous field; then, we have (see also Exercise 14.8 in Ch. 14)

$$P_N = \frac{1}{\bar{n}_{\rm g}}. \tag{11.24}$$

While simple Poisson sampling is not a realistic assumption for actual galaxies for a whole host of reasons, a scale-independent noise $P_N$ is nevertheless expected at low $k$.

Now, if we measure $P_{\rm g,RSD}(k,\mu_k)$, we can vary both $k$ and $\mu_k$, allowing us to disentangle the contributions multiplied by $b_1$ and $f$. Technically, this is often done by performing a multipole decomposition of $P_{\rm g,RSD}(k,\mu_k)$ with respect to $\mu_k$, see Exercise 11.4. Fig. 11.4 shows the recent measurements of the galaxy power spectrum multipoles from the BOSS survey (part of SDSS-III). The measurements show strikingly small error bars and good agreement with the model, which consists of the linear prediction derived here as well as nonlinear corrections which we discuss in Ch. 12. Notice that the error bars increase toward low $k$ (large scales). The reason is the same sample variance that leads to larger errors for the low-$l$ CMB multipoles: only a finite number of Fourier modes are probed in a finite survey volume $V_{\rm survey}$. The number is approximately $N_k = 2\pi k^2 \Delta k V_{\rm survey}/(2\pi)^3$. As long as $V_{\rm survey}$ does not contain the entire observable universe, we can thus measure more modes by performing larger and deeper galaxy surveys (i.e. by devoting more observing time and larger telescopes to the task).

Do our results mean that we can measure both the galaxy bias and the growth rate? Not quite: in general, we do not know the matter power spectrum $P_{\rm L}(k,\bar{z})$. In particular, its amplitude is not directly measurable. But all is not lost: if we denote the amplitude of the matter power spectrum with $\sigma_8$ as defined in Exercise 8.13, then the three-dimensional

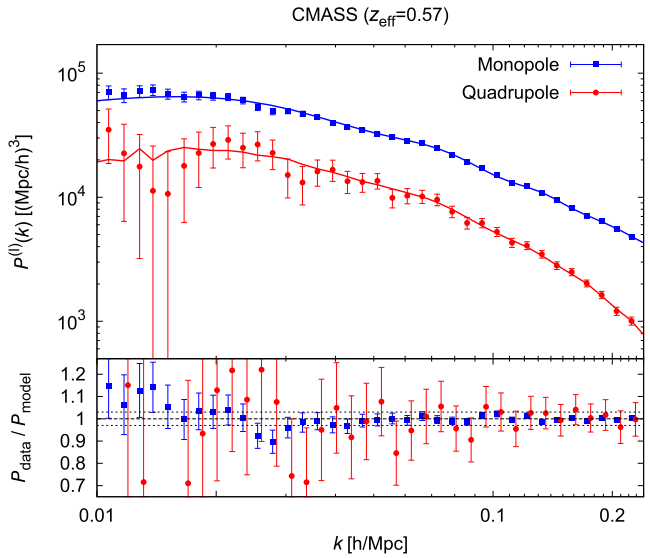



**FIGURE 11.4** Three-dimensional galaxy power spectrum of the CMASS sample observed by the BOSS survey. Shown are the monopole and quadrupole moments with respect to the cosine $\mu_k$ of the wavevector with the line of sight. The bottom panel shows the ratio of the data to the best-fitting model, which includes Eq. (11.23) as well as nonlinear corrections at higher $k$. From Gil-Marín et al. (2016).

galaxy power spectrum allows us to measure $b_1\sigma_8$ and $f\sigma_8$. The former quantity is galaxy-dependent and more or less a nuisance parameter for most cosmologists. $f\sigma_8$ on the other hand contains valuable information.

Constraints on $f\sigma_8$ are summarized in Fig. 11.5. As we have seen in Sect. 8.5, the growth factor and hence growth rate are a direct probe of dark energy. The faster the expansion due to dark energy, the lower the growth rate. Conversely, should gravity be in fact not described by general relativity, we expect a larger growth rate, since modifications to gravity typically increase the strength of gravity. An especially important test thus is to compare the measured growth rate with that expected for the measured expansion rate; general relativity predicts a unique relation regardless of the type of dark energy (Eq. (8.78)), while modified gravity changes this relation. Remarkably, the data do support the Euclidean concordance cosmology with $\Omega_\Lambda \simeq 0.7$ (solid line): the high-redshift points show the trend expected for a matter-dominated universe, while there is evidence for a suppression relative to that trend, indeed a turnover, at redshifts $z \lesssim 0.5$.

### 11.1.3 BAO and Alcock–Paczyński

Let us now turn to the effect of a "wrong cosmology:" That is, we use a different distance-redshift relation to assign the 3D galaxy positions via Eq. (11.1) than the one of the true universe. Unless some higher power tells us the truth about our universe, this will certainly

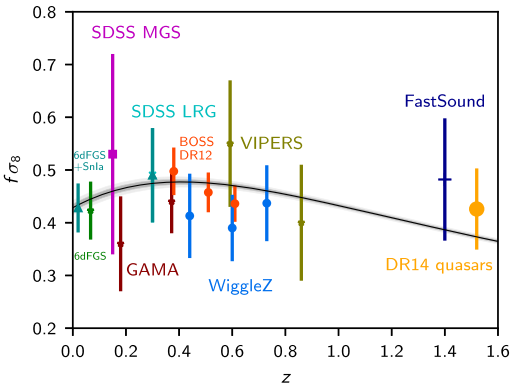



**FIGURE 11.5** Constraints on the parameter $f\sigma_8$ placed from redshift-space distortions measured in different surveys, as well as direct measurements of velocities (the two lowest-$z$ data points). Direct measurements use distance indicators such as supernovae to estimate the peculiar velocity from the observed redshift and estimated distance, via Eq. (11.4). From Planck Collaboration (2018b).

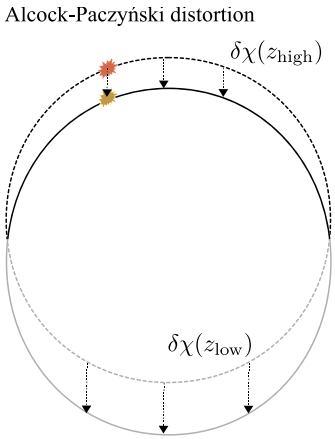



**FIGURE 11.6** Illustration of the Alcock–Pacyński distortion due to a different cosmology used in the assignment of 3D galaxy positions. To lowest order, all galaxies are displaced along the line of sight (dashed arrows) from the true positions (dashed circle) by the same amount $\delta\chi(\bar{z})$. However, since $\delta\chi$ in general evolves with redshift, the displacement differs slightly between galaxies that are more nearby ($z_{\rm low}$; bottom half) and those further away ($z_{\rm high}$; top half). The result is the solid ellipse which indicates the apparent locus of the galaxies.

be the case! Fortunately, the observed galaxy power spectrum can help tell us what the true distance-redshift relation is.

This is in fact somewhat easier to derive than redshift-space distortions; we have already done most of the work. The basic effect is illustrated in Fig. 11.6. The galaxies are displaced from their true position by an amount $\delta\chi(z)$. At lowest order, this is the same for all galaxies, but since the distance-redshift relation evolves differently in different cos-

mologies, the displacement is different for galaxies at different cosmological redshifts, leading to a distortion of the dashed circle in Fig. 11.6 into an ellipse.

As observers, we turn the observed galaxy position $(\theta, \phi)$ and redshift $z$ into an "observed" 3D position $\boldsymbol{x}_{\rm obs}$. We do this by using a fiducial distance-redshift relation $\chi_{\rm fid}(z)$, and by defining a convenient origin of the 3D coordinate system. Let us again make use of the flat-sky approximation, so that the position on the sky becomes a 2D vector $\boldsymbol{\theta}$, and we choose the origin such that

$$\boldsymbol{x}_{\rm obs} = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{\theta} = \mathbf{0},\ z = \bar{z}, \tag{11.25}$$

where $\boldsymbol{\theta} = 0$ corresponds to a point on the sky near the center of the survey footprint, and $\bar{z}$ is the central value of the redshift slice considered. Now, the transverse components $x^1_{\rm obs}$, $x^2_{\rm obs}$ we *assign* to the galaxy are

$$(x^1_{\rm obs}, x^2_{\rm obs}) = \chi_{\rm fid}(z) \times (\theta^1, \theta^2)\,, \tag{11.26}$$

while the components we *should assign* in the true cosmology are

$$(x^1, x^2) = \chi(z) \times (\theta^1, \theta^2) = \left[1 - \frac{\delta\chi(z)}{\chi_{\rm fid}(z)}\right](x^1_{\rm obs}, x^2_{\rm obs})\,. \tag{11.27}$$

The second equality is obtained by subtracting Eq. (11.26) and through Eq. (11.2). Throughout we will work to linear order in $\delta\chi$. Eq. (11.27) comes as no surprise: if $\delta\chi > 0$, then we assign galaxies a larger comoving distance than they actually have. Correspondingly, the true galaxy position is closer to the origin in the transverse directions ($|x^1|$, $|x^2|$ are smaller) than the assigned position. The opposite holds if $\delta\chi < 0$.

The line-of-sight component of the position $x^3_{\rm obs}$ is determined by the redshift. Since we have chosen $z = \bar{z}$ to correspond to $x^3 = 0$, the observed coordinate is

$$x^3_{\rm obs}(z) = \chi_{\rm fid}(z) - \chi_{\rm fid}(\bar{z}) \simeq \frac{1}{H_{\rm fid}(\bar{z})}(z - \bar{z}), \tag{11.28}$$

where we have expanded to linear order in $z - \bar{z}$, under the assumption that we are considering a narrow redshift slice so that this difference is always small, and used $d\chi/dz = 1/H$. The same relation holds for the line-of-sight component $x^3(z)$ that we *should* assign to the galaxy:

$$x^3(z) \simeq \frac{1}{H(\bar{z})}(z - \bar{z}) = \frac{H_{\rm fid}(\bar{z})}{H(\bar{z})} x^3_{\rm obs}. \tag{11.29}$$

Using $\delta H(z) = H(z) - H_{\rm fid}(z)$, and working to linear order in $\delta H$, we can rewrite the second equality as

$$x^3(z) = \left[1 - \frac{\delta H(\bar{z})}{H_{\rm fid}(\bar{z})}\right] x^3_{\rm obs}. \tag{11.30}$$

Since the zero-point of the 3-axis is defined by a fixed redshift $\bar{z}$, the displacement between true and assigned positions is only induced by a difference in how $\chi(z)$ and $\chi_{\rm fid}(z)$ vary around $\bar{z}$; that is, the relevant quantity is the slope $d\chi/dz$ of the distance-redshift relation at $\bar{z}$, which is the inverse of the Hubble parameter $H(\bar{z})$. We see that the line-of-sight displacement of the position (Eq. (11.30)) is different from the transverse displacement (Eq. (11.27)), leading in general to an elliptical distortion of the galaxy density field (Fig. 11.6).

To summarize, the relation between the true and observed galaxy positions due to a wrong distance-redshift relation is

$$\begin{aligned} \boldsymbol{x}(\boldsymbol{x}_{\rm obs}) &= \left([1-\alpha_\perp]x^1_{\rm obs},\ [1-\alpha_\perp]x^2_{\rm obs},\ [1-\alpha_\parallel]x^3_{\rm obs}\right), \quad \text{where} \\ \alpha_\perp &= \left.\frac{\delta\chi}{\chi_{\rm fid}}\right|_{\bar{z}}; \quad \alpha_\parallel = \left.\frac{\delta H}{H_{\rm fid}}\right|_{\bar{z}}. \end{aligned} \tag{11.31}$$

Again, since the galaxies are in a narrow redshift slice, evaluating $\alpha_\perp$, $\alpha_\parallel$ at the mean redshift $\bar{z}$ is sufficient.

We will soon see that by measuring the observed power spectrum, we can obtain constraints on $\alpha_\perp$ and $\alpha_\parallel$. First, though, let us consider what we could learn from such a measurement. Given a measurement of $\alpha_\perp$, the first part of the second line of Eq. (11.31) can be used to express $\chi(\bar{z})$ in terms of known quantities:

$$\chi(\bar{z}) = \chi_{\rm fid}(\bar{z})\ [1+\alpha_\perp] \tag{11.32}$$

and similarly $\alpha_\parallel$ yields

$$H(\bar{z}) = H_{\rm fid}(\bar{z})\ \left[1+\alpha_\parallel\right]. \tag{11.33}$$

That is, a measurement of these two distortion parameters enables us to infer the distance ($\chi(\bar{z})$) and the Hubble rate ($H(\bar{z})$) at a given redshift $\bar{z}$.

We now show how $\alpha_\perp, \alpha_\parallel$ can be extracted from the galaxy power spectrum. This follows in analogy to Eq. (11.20). Having computed the contribution from redshift-space distortions in the previous section, we now only have to include the effect of the coordinate rescaling Eq. (11.31). In this case, we have to keep track of the effect on the argument of $\delta_{\rm g}$, since the rescaling is spatially uniform, and hence corresponds to a zeroth-order effect (unlike the case in RSD, where the displacement due to peculiar velocities was first order). Finally, we should now include the factor $\bar{J}$ of the Jacobian which we have dropped after Eq. (11.15), since it will ensure the correct normalization. We obtain

$$\begin{aligned} \delta_{\rm g,obs}(\boldsymbol{k}_{\rm obs}) &= \bar{J}\int d^3x_{\rm obs}\ e^{-i\boldsymbol{k}_{\rm obs}\cdot\boldsymbol{x}_{\rm obs}}\delta_{\rm g,RSD}(\boldsymbol{x}[\boldsymbol{x}_{\rm obs}]) \\ &= \bar{J}(1+\alpha_\perp)^2(1+\alpha_\parallel)\int d^3x\ e^{-i\boldsymbol{k}[\boldsymbol{k}_{\rm obs}]\cdot\boldsymbol{x}}\delta_{\rm g,RSD}(\boldsymbol{x}) \\ &= \delta_{\rm g,RSD}(\boldsymbol{k}[\boldsymbol{k}_{\rm obs}]), \end{aligned} \tag{11.34}$$

where we have inverted Eq. (11.31) in the second line, and defined

$$\boldsymbol{k}[\boldsymbol{k}_{\rm obs}] = \left([1+\alpha_\perp]k^1_{\rm obs},\ [1+\alpha_\perp]k^2_{\rm obs},\ [1+\alpha_\parallel]k^3_{\rm obs}\right). \tag{11.35}$$

The prefactors in Eq. (11.34) simply reduce to unity via Eq. (11.14). This is not a miracle, but follows from the fact that the number of galaxies in a given volume is independent of the coordinates that we use to describe them, so $N_{\rm g} = n_{\rm g}\Delta x^3 = n_{\rm g,obs}\Delta x^3_{\rm obs}$. The integral over $d^3x$ yields $\delta_{\rm g,RSD}$, i.e. Eq. (11.22), evaluated at $\boldsymbol{k}[\boldsymbol{k}_{\rm obs}]$. That is, the difference between fiducial and true cosmologies manifests itself in the Fourier-space galaxy density through a simple rescaling of the wavevector:

$$\delta_{\rm g,obs}(\boldsymbol{k}_{\rm obs}) = [b_1 + f\mu_k^2]\delta_{\rm m}(\boldsymbol{k})\Big|_{\boldsymbol{k}=([1+\alpha_\perp]k^1_{\rm obs},\ [1+\alpha_\perp]k^2_{\rm obs},\ [1+\alpha_\parallel]k^3_{\rm obs})}. \tag{11.36}$$

Finally, Eq. (11.23) can now be used to write the observed galaxy power spectrum as

$$P_{\rm g,obs}(\boldsymbol{k}_{\rm obs},\bar{z}) = \left(P_{\rm L}(k,\bar{z})\left[b_1 + f\mu_k^2\right]^2\right)\Big|_{\boldsymbol{k}=([1+\alpha_\perp]k^1_{\rm obs},\ [1+\alpha_\perp]k^2_{\rm obs},\ [1+\alpha_\parallel]k^3_{\rm obs})} + P_N. \tag{11.37}$$

This equation includes two effects: the redshift-space distortions due to peculiar velocities derived in the previous section; and the fact that the coordinates we assign to galaxies are based on an assumed distance-redshift relation, not the true one. The constant noise term $P_N$ is unaffected by both of these.

Note that, even if there were no galaxy velocities, an incorrect distance-redshift relation would lead to an anisotropy in the galaxy power spectrum: imagine setting $f = 0$, which is formally equivalent to setting all velocities to zero. The relation between $\boldsymbol{k}_{\rm obs}$ and $\boldsymbol{k}$ now still depends on the angle of $\boldsymbol{k}_{\rm obs}$ with the line of sight, because $\alpha_\parallel$ is different from $\alpha_\perp$. Thus, a wrong assumed cosmology induces an anisotropy in the galaxy power spectrum. This fact was first pointed out by Alcock and Paczyński (1979), and is thus known as *Alcock–Paczyński (AP)* effect. Unlike RSD, the amplitude of the AP effect depends on the shape of the power spectrum, as you can see by expanding Eq. (11.37) to linear order in $\alpha_\parallel$, $\alpha_\perp$. For this reason, AP and RSD can be disentangled.

Finally, Eq. (11.37) is also the basis of one of the prime science targets of current and upcoming galaxy redshift surveys: using the baryon acoustic oscillation feature as a standard ruler, usually simply abbreviated as "BAO" (Fig. 11.7). Recall that the matter power spectrum contains a small, oscillatory modulation (see Sect. 8.6.1), roughly of the form $\cos(kr_s)$, where $r_s \approx 105\,h^{-1}$ Mpc is the sound horizon at recombination. In the early universe, this feature was imprinted only in the baryonic component of matter, but since baryons and dark matter are coupled by gravity, this oscillatory pattern is transferred to the late-time power spectrum of matter, albeit with smaller amplitude.

This means that we have a well-defined feature in the power spectrum at a *true* comoving scale $k \sim \pi/r_s$. Eq. (11.37) says that we will observe the same feature in the galaxy power spectrum, but at an apparent scale $k_{\rm obs}[k]$. Since $r_s$ is extremely well determined

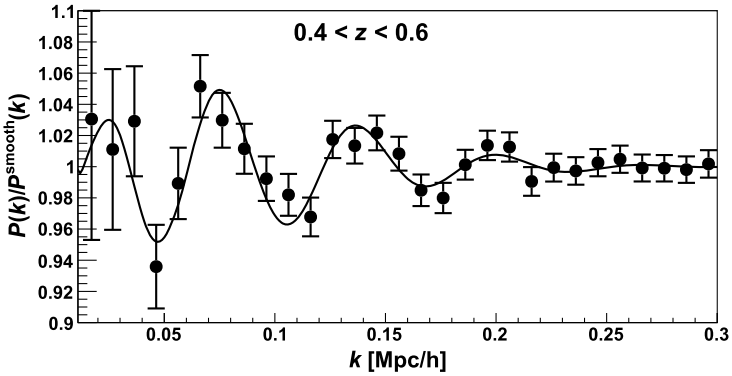



**FIGURE 11.7** The BAO feature in the angle-averaged (monopole) three-dimensional galaxy power spectrum of the CMASS sample observed by the BOSS survey. Both data (points) and best-fit model (line) have been divided by a smooth model power spectrum without the BAO feature, in order to enhance the visibility of the feature. The fit parameters include $\alpha_\perp$ and $\alpha_\parallel$, which shift the model along the $k$-axis [see Eq. (11.37)]. Adapted from Beutler et al. (2017).

by the CMB,[1] measuring this feature in the galaxy power spectrum at redshift $\bar{z}$ allows us to measure $\alpha_\perp$ and $\alpha_\parallel$ precisely (see Fig. 11.7). Eq. (11.37) describes rigorously what is much simpler to understand intuitively: the BAO oscillations are a feature of a known size imprinted in the clustering pattern of galaxies. Measuring its apparent size observed in galaxies at a redshift $\bar{z}$ allows us to measure the distance to that redshift. In particular, the BAO feature gives us a direct measurement of the distance[2] $\chi(\bar{z})$ (via $\alpha_\perp$) and of the Hubble rate $H(\bar{z})$ (via $\alpha_\parallel$). Because of the well-understood theoretical underpinning and characteristic shape of the BAO feature, this approach of measuring distances is extremely robust.

To summarize this section, the large-scale three-dimensional galaxy power spectrum as measured in galaxy redshift surveys contains two important sources of cosmological information:

- RSD yield the growth rate $f\sigma_8(\bar{z})$ via the amplitude and anisotropy of the power spectrum.
- The BAO feature and AP distortions allow us to measure $d_A(\bar{z})$ and $H(\bar{z})$, by using the difference in the fiducial and true distance-redshift relations.

No wonder then that significant resources are being devoted to increasingly larger redshift surveys: they contain a rich amount of information both on the background expansion history and the growth of structure. Moreover, all of what we described applies to any tracer

[1] In fact, since the sound horizon mainly depends on the baryon density $\Omega_b h^2$, the constraint from BBN combined with Deuterium abundance suffices as well.

[2] More precisely the angular diameter distance $d_A(\bar{z})$, which is equivalent in our case since we have assumed a Euclidean universe throughout.

of large-scale structure, not just galaxies: for example, quasars, the Lyman-alpha forest, and unresolved line emitters and absorbers (known as intensity mapping).

We should emphasize again though that this information relies on our assumptions of a simple linear relation between the tracer density and matter density perturbations (Eq. (11.18)), as well as unbiased tracer velocities. After all, if the latter were not the case, we could not infer an unbiased estimate of the growth rate; if bias was more complicated, we could not be sure that the BAO feature in the tracer power spectrum is at the same scale as in the matter power spectrum. These assumptions thus need to be carefully justified. In Ch. 12, we will do precisely that.

## 11.2 Angular correlations

In the previous section, we derived the three-dimensional galaxy power spectrum measured from a set of galaxy positions and precise redshifts. But what if we do not have those redshifts? Large imaging surveys identify the sky positions of many millions of galaxies, but without expensive followup observations, their distances remain uncertain. Can we still extract information from such a data set? The answer is yes, and it is in fact not much more difficult. In many cases, surveys observe in several bands, so that colors are available for the galaxies. These colors can be turned into approximate redshift estimates, so-called photometric redshifts, which can be used as proxies (with significant scatter) for the true redshifts.

So when dealing with imaging surveys, we do not have individual distances of galaxies but we have a handle on the *distribution* of distances, $W(\chi)$:

$$W(\chi) = \frac{1}{N_\mathrm{g}} \frac{dN_\mathrm{g}}{d\chi}, \tag{11.38}$$

where $N_\mathrm{g}$ is the total number of galaxies, and $W(\chi)$ is normalized to unity over the interval $\chi \in [0, \infty)$. In practice, $W(\chi)$ drops to zero below and above some minimum and maximum distances. Galaxies at large distances are too faint to be detected, and there are not that many galaxies at low redshifts simply because the volume is small. Photometric redshifts are notoriously difficult, so determining $W(\chi)$ is often a daunting task on its own. We will assume here that this has been taken care of.

Instead of measuring the 3D galaxy density field, we now measure its projection on the sky. Practically, we can imagine dividing the sky area covered by the survey into many small pixels, and counting the galaxies in each pixel. Subtracting and dividing by the mean, we obtain the projected overdensity $\Delta_\mathrm{g}(\hat{\boldsymbol{n}})$. This is just a superposition of many slices of the 3D galaxy density field at different distances $\chi$, weighted by the distance distribution, so we have

$$\Delta_\mathrm{g}(\hat{\boldsymbol{n}}) = \int_0^\infty d\chi\, W(\chi) \delta_{\mathrm{g,obs}} \left(\boldsymbol{x} = \hat{\boldsymbol{n}}\chi, \eta = \eta_0 - \chi\right). \tag{11.39}$$

We use $\Delta_{\rm g}$ to distinguish the projected galaxy density from the three-dimensional one, $\delta_{\rm g,obs}$. Notice that the projection involves the galaxy density at different times $\eta$, since the photons all travel at the speed of light, so that the more distant galaxies are seen at an earlier time. Let us now insert the Fourier transform of $\delta_{\rm g,obs}$, and use the expansion of the exponential in Eq. (C.17):

$$\begin{aligned}\Delta_{\rm g}(\hat{\boldsymbol{n}}) &= \int_0^\infty d\chi\, W(\chi) \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\hat{\boldsymbol{n}}\chi}\, \delta_{\rm g,obs}(\boldsymbol{k},\eta(\chi)) \\ &= 4\pi \int \frac{d^3k}{(2\pi)^3} \sum_{lm} i^l Y_{lm}(\hat{\boldsymbol{n}}) Y^*_{lm}(\hat{\boldsymbol{k}}) \int_0^\infty d\chi\, W(\chi)\, j_l(k\chi)\, \delta_{\rm g,obs}(\boldsymbol{k},\eta(\chi)) \end{aligned} \tag{11.40}$$

where we have abbreviated $\eta(\chi) = \eta_0 - \chi$, and $\sum_{lm} \equiv \sum_{l=0}^\infty \sum_{m=-l}^l$. The right-hand side is nothing but an expansion of $\Delta_{\rm g}(\hat{\boldsymbol{n}})$ in spherical harmonics, which we can read off as the coefficients of $Y_{lm}(\hat{\boldsymbol{n}})$:

$$\Delta_{{\rm g},lm} = 4\pi i^l \int \frac{d^3k}{(2\pi)^3} Y^*_{lm}(\hat{\boldsymbol{k}}) \int_0^\infty d\chi\, W(\chi)\, j_l(k\chi)\, \delta_{\rm g,obs}(\boldsymbol{k},\eta(\chi)). \tag{11.41}$$

In exact analogy with the CMB anisotropies (the $a_{lm}$), the angular power spectrum of galaxy counts on the sky is then proportional to the expectation value of $|\Delta_{{\rm g},lm}|^2$. Let us thus evaluate

$$\begin{aligned}\left\langle \Delta_{{\rm g},lm} \Delta^*_{{\rm g},l'm'} \right\rangle &= (4\pi)^2 i^{l-l'} \int \frac{d^3k}{(2\pi)^3} \int \frac{d^3k'}{(2\pi)^3} Y^*_{lm}(\hat{\boldsymbol{k}}) Y_{l'm'}(\hat{\boldsymbol{k}}') \int_0^\infty d\chi\, W(\chi)\, j_l(k\chi) \\ &\quad \times \int_0^\infty d\chi'\, W(\chi')\, j_{l'}(k'\chi') \left\langle \delta_{\rm g,obs}(\boldsymbol{k},\eta(\chi)) \delta^*_{\rm g,obs}(\boldsymbol{k}',\eta(\chi')) \right\rangle. \end{aligned} \tag{11.42}$$

The brackets $\langle \ldots \rangle$ here denote an ensemble average over all realizations of the density field. The ensemble average over the two fields immediately sets $\boldsymbol{k}' = \boldsymbol{k}$ (due to homogeneity), and we can use the orthonormality of spherical harmonics (Eq. (C.11)) to obtain

$$\left\langle \Delta_{{\rm g},lm} \Delta^*_{{\rm g},l'm'} \right\rangle = \delta_{ll'} \delta_{mm'} C_{\rm g}(l) \tag{11.43}$$

where the angular power spectrum is defined as

$$\begin{aligned} C_{\rm g}(l) &= \frac{2}{\pi} \int k^2 dk \int_0^\infty d\chi\, W(\chi)\, j_l(k\chi) \int_0^\infty d\chi'\, W(\chi')\, j_l(k\chi') \\ &\quad \times P_{\rm g,obs}(\boldsymbol{k},\eta(\chi),\eta(\chi')). \end{aligned} \tag{11.44}$$

Notice that the angular power spectrum $C_{\rm g}(l)$ of galaxies in general involves the *unequal-time* power spectrum of galaxies, since we are projecting along the lightcone. This unequal-time power spectrum is nonzero, because the density perturbations remain in place as they grow. We will see very soon though that on small scales, i.e. large $l$, only equal times and distances $\chi' = \chi$ contribute appreciably.

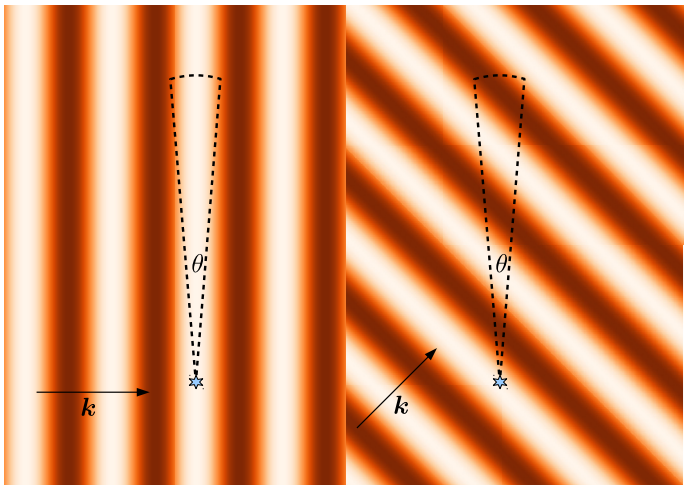



**FIGURE 11.8** Two plane-wave perturbations and their contributions to the angular power spectrum. Right panel shows a perturbation with longitudinal wavenumber $\mu_k k \gg \chi^{-1}$ (the $\hat{\boldsymbol{e}}_z$-direction is vertical). Left panel shows an essentially transverse mode with $\mu_k k < \chi^{-1}$. Angular correlations due to the mode in the right panel are negligible since there are cancelations along the line of sight, while those do not occur for the transverse mode shown on the left.

Eq. (11.44) is the exact result for the angular power spectrum of galaxies, given their three-dimensional power spectrum $P_{\rm g,obs}(\boldsymbol{k}, \eta, \eta')$ (allowing for anisotropy) and the selection function $W(\chi)$. However, it is cumbersome to handle since it involves three nested integrals, and these integrals are over the oscillating functions $j_l$. Moreover, we need to specify the full, unequal-time galaxy power spectrum.

In order to proceed, we make a simplifying assumption analogous to the distant-observer approximation employed in the previous section. On small scales, $l \gg 1$, the galaxy pairs contributing to $C_{\rm g}(l)$ subtend a small angle on the sky, roughly $\theta \sim 1/l$, so we expect some simplifications. Let us then look more closely at Eq. (11.44) in this regime. The integral over $k$ is

$$\frac{2}{\pi}\int k^2 dk\, j_l(k\chi)\, j_l(k\chi')\, P_{\rm g,obs}(\boldsymbol{k}, \eta, \eta'). \tag{11.45}$$

In Exercise 11.6, you show that, if $P_{\rm g,obs}(\boldsymbol{k})$ were *independent of* $k$, so that it can be pulled out of the integral, this integral would reduce to

$$\frac{2}{\pi}\int k^2 dk\, j_l(k\chi)\, j_l(k\chi') = \frac{1}{\chi^2}\delta_{\rm D}^{(1)}(\chi - \chi'). \tag{11.46}$$

With this, Eq. (11.44) becomes much simpler; it reduces to a single integral over $\chi$. But in reality, $P_{\rm g,obs}$ is *not* independent of $k$ of course. Let us then inspect the remainder of the integrand in Eq. (11.45). As you can show in Exercise 11.7, for high $l$, the product of spherical Bessel functions is very sharply peaked at $k\chi \approx k\chi' \approx \sqrt{l(l+1)} \approx l + 1/2$. As long

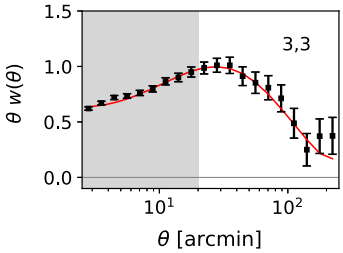



**FIGURE 11.9** Angular correlation function of galaxies in the photometric Dark Energy Survey (the "3,3" indicates autocorrelation of galaxies within a photometric redshift range centered at $z \simeq 0.55$). The correlation function $w(\theta) = w_{\rm g}(\theta)$ has been multiplied by $\theta$ in order to reduce the dynamic range, as $w_{\rm g}$ grows strongly toward small $\theta$. The gray shaded region involves comoving scales smaller than $8\,h^{-1}\,{\rm Mpc}$, which are significantly affected by nonlinear evolution and bias. The line shows the best-fit model based on the linear bias prescription. From Elvin-Poole et al. (2018).

as $P_{\rm g,obs}(\boldsymbol{k})$ varies slowly over the narrow range $\Delta k$ over which the Bessel functions are nonzero, $\Delta k \sim 1/(l\chi)$, we can approximate it as constant. The approximation we have just described, which is usually very accurate at $l \gtrsim 20$, is known as the *Limber approximation*. Its core prediction then is

$$C_{\rm g}(l) = \int \frac{d\chi}{\chi^2} W^2(\chi) P_{\rm g,obs}\left(k = \frac{l+1/2}{\chi},\, \mu_k = 0,\, \eta(\chi)\right), \tag{11.47}$$

which is much faster to calculate than Eq. (11.44). The remaining arguments in $P_{\rm g,obs}$ deserve some explanation: we have seen that $\chi = \chi'$ in the Limber approximation, which implies $\eta(\chi') = \eta(\chi)$, so that Eq. (11.47) only involves the equal-time power spectrum. It also means that the $k$ modes involved do not have a line-of-sight component, since that would mean different distances of different points along the perturbation, i.e. $\chi' \neq \chi$. So, $\boldsymbol{k}$ has to be transverse to the line of sight: $\mu_k = 0$.

Our derivation so far was quite rigorous, but mathematical. Fig. 11.8 illustrates the physical reason behind the Limber approximation. Focusing on small scales corresponds to looking at small angles, $\theta \sim 1/l \ll 1$. Now consider the figure. Modes with longitudinal wavenumber $\mu_k k$ much greater than $\chi^{-1}$ do not give rise to angular correlations because of cancelations along the line of sight. Only modes with $\mu_k k$ of order $\chi^{-1}$ or smaller lead to angular correlations. Therefore, the relevant transverse wavenumbers $l/\chi$ are much larger than the relevant longitudinal wavenumbers, and we can safely neglect the latter. This then corresponds to setting $\chi' = \chi$.

Finally, we can also write down the angular correlation function $w_{\rm g}(\theta)$. The full-sky relation between $C_{\rm g}(l)$ and $w_{\rm g}(\theta)$ is derived in Exercise 11.8. On small scales, however, in the flat-sky approximation, we can treat $C_{\rm g}(l)$ as the 2D power spectrum on a plane, so

$$w_{\rm g}(\theta) = \int \frac{d^2 l}{(2\pi)^2} e^{i\boldsymbol{l}\cdot\boldsymbol{\theta}} C_{\rm g}(l). \tag{11.48}$$

Since $C_g(l)$ depends only on the magnitude of $\boldsymbol{l}$, the angular part of the integration over $l$ is $\int_0^{2\pi} d\phi\ e^{il\theta\cos\phi}$, which is proportional to $J_0(l\theta)$, the Bessel function of order zero (Eq. (C.24)). Therefore,

$$w_g(\theta) = \int_0^\infty \frac{dl}{2\pi} l C_g(l) J_0(l\theta). \tag{11.49}$$

Fig. 11.9 shows the projected correlation function measured by the Dark Energy Survey (DES). The measurement has very high signal-to-noise, thanks to the many galaxies that are available. Unfortunately, due to the projection within a wide redshift slice, the BAO feature is smoothed out and very difficult to detect. Nevertheless, the combination of this projected correlation function with weak lensing allows for many cosmological tests. We turn to this in Ch. 13.

## 11.3 The Sunyaev–Zel'dovich effect

The most obvious tracers of large-scale structure are objects at low redshifts such as galaxies. But large-scale structure also leads to characteristic imprints on the CMB, whose photons travel through the entire observable universe to reach us. Using the CMB as a backlight allows us to probe large-scale structure at fairly high redshifts, at which point a direct observation of luminous tracers becomes increasingly difficult. We have already encountered one such imprint in Sect. 9.6: the integrated Sachs–Wolfe effect, which is only relevant on very large scales. Another effect is *CMB lensing*, the deflection of CMB photon trajectories by the gravitational potential of large-scale structure, which we will turn to in Ch. 13. In this section, we study another important effect, the scattering of CMB photons off of ionized gas in the late universe.

As we learned in Ch. 4 and Ch. 9, most of the gas in the universe at redshifts below $z \sim 6$ is ionized. This means that CMB photons are no longer completely decoupled from the gas, but they are able to scatter (Fig. 11.10; notice that unlike the other two effects mentioned above, here the CMB photons can dramatically change direction). This scattering is fortunately much less efficient than it was prior to recombination, due to the much lower density of the gas, leading to an optical depth much less than 1. In Ch. 9 we saw that the scattering leads to a damping of anisotropies in the CMB, and in Ch. 10 that it sources polarization on very large angular scales.

Apart from the much lower gas density, another significant difference in this scattering process compared to that happening before recombination is that the gas is now much hotter than the CMB photons: while the CMB temperature is 20 K and falling at $z \leq 6$, the gas temperature ranges from $10^4$ K to temperatures in excess of $10^7$ K in massive clusters. This means that electrons are much more energetic than the CMB photons, so that they tend to *increase* the energy of CMB photons in the scattering process (an effect known as inverse-Compton scattering). Therefore, after some fraction of photons have been scattered thereby gaining energy, the CMB spectrum is distorted and no longer is a perfect black-body. By observing the CMB at different frequencies, this distortion of the spectrum

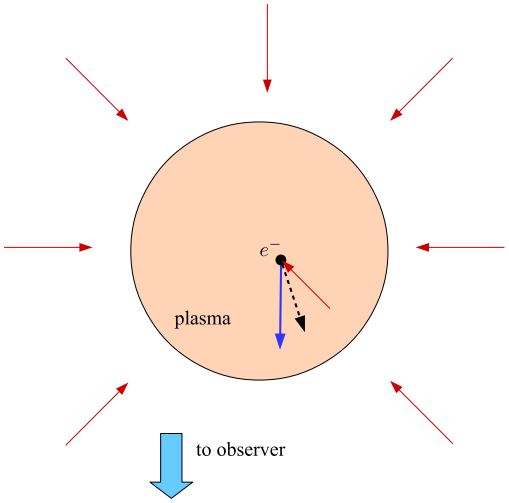



**FIGURE 11.10** Compton scattering of CMB photons (thin red arrows) within a cloud (circle) of hot ionized gas. A fraction of the CMB photons are scattered off electrons in the cloud (black dot), and reach the observer typically with a small energy gain (blue thick arrow). Notice that the typical momentum $q \sim \sqrt{m_e T_e}$ of electrons in the plasma is much larger than that of the CMB photons $p \sim T$.

can be distinguished from the CMB temperature perturbations that we have studied so far in this book. This upscattering *SZ effect* was first pointed out by Zel'dovich and Sunyaev (1969).

Our starting point is, not surprisingly, the Boltzmann equation for photons, with a collision term due to Compton scattering. We thus start from

$$\left[\frac{\partial}{\partial t} - Hp\frac{\partial}{\partial p}\right] f(p,t) = C[f(p)], \tag{11.50}$$

with the collision term as derived in Eq. (5.13):

$$C[f(\boldsymbol{p})] = \frac{\pi}{2m_e p}\int \frac{d^3q}{(2\pi)^3 2m_e}\int \frac{d^3p'}{(2\pi)^3 2p'}\delta_{\rm D}^{(1)}\left[p + \frac{q^2}{2m_e} - p' - \frac{(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')^2}{2m_e}\right] \times \sum_{3\ \text{spins}} |\mathcal{M}|^2 \{f_e(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')f(\boldsymbol{p}') - f_e(\boldsymbol{q})f(\boldsymbol{p})\}. \tag{11.51}$$

Here, we continue to make several assumptions made in Ch. 5: the non-relativistic regime, $T$, $T_e \ll m_e$, where $T$ is the CMB temperature at the redshift where the scattering takes place, while $T_e$ is the electron temperature; isotropic scattering, taking $|\mathcal{M}|^2$ as well as the distribution functions to be independent of $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{q}}$; and that quantum effects such as stimulated emission are negligible. All of these are similarly or more accurate in this application as before recombination, except $T_e \ll m_e$: in very massive galaxy clusters, this assumption ceases to be accurate, leading to minor modifications of the resulting photon energy spec-

trum. We also neglect the gravitational potential term on the left-hand side of Eq. (11.50), since it does not affect the spectral distortion we are interested in.

In Ch. 5, we proceeded to expand the Dirac delta to linear order in the energy shift of the photons $p - p'$ (Eq. (5.15)). This resulted in the coupling of photons to the baryon velocity, but no change in the photon spectrum; that is, we could still phrase the effect in terms of a perturbation $\Theta(\hat{\boldsymbol{p}})$ to the photon temperature. Now, let us examine the *second-order* piece in this expansion:

$$\delta_{\rm D}^{(1)}\left[p + \frac{q^2}{2m_e} - p' - \frac{(\boldsymbol{q}+\boldsymbol{p}-\boldsymbol{p}')^2}{2m_e}\right] = \text{(Eq. (5.15))} + \frac{1}{2}\left[\frac{(\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{q}}{m_e}\right]^2 \frac{\partial^2}{\partial p'^2}\delta_{\rm D}^{(1)}(p-p'). \tag{11.52}$$

Consider now the contribution of this second-order term (in the momentum transfer) to the collision integral. This is analogous to the steps that led us to Eq. (5.19), and you will perform this derivation in Exercise 11.9. Choosing the $z$-direction to lie along the direction of the electron momentum, we arrive at the following contribution to the collision term

$$C[f(p)]\Big|_{\rm SZ} = 2\pi^2 \frac{\sigma_{\rm T}}{m_e}\int \frac{d^3q}{(2\pi)^3} f_e(q)\frac{q^2}{m_e} \times \int \frac{d\Omega'}{(2\pi)^3}\int dp' \delta_{\rm D}^{(1)}(p-p')\frac{\partial^2}{\partial p'^2}\left[p'(p_z - p_z')^2(f(p)-f(p'))\right]. \tag{11.53}$$

We recognize the integral over the electron distribution function as the kinetic energy density (taking into account the degeneracy factor $g_e = 2$ for electrons), which for a gas (more precisely, plasma) of temperature $T_e$ is $3n_eT_e/2$. The remaining steps are now straightforward (Exercise 11.9). After averaging over $\mu = p_z/p$, i.e. integrating $\int_{-1}^{1} d\mu/2$, we finally obtain

$$C[f(p)]\Big|_{\rm SZ} = \frac{n_eT_e\sigma_{\rm T}}{m_e}\left[4p\frac{\partial f}{\partial p} + p^2\frac{\partial^2 f}{\partial p^2}\right] = \frac{n_eT_e\sigma_{\rm T}}{m_e}\frac{1}{p^2}\frac{\partial}{\partial p}\left[p^4\frac{\partial f}{\partial p}\right]. \tag{11.54}$$

In Exercise 11.10, you show that the integral over all $p$ of this collision term vanishes. That is, the number of photons is conserved. This comes as no surprise, since Compton scattering conserves the photon number.

Let us now replace the momentum $p$ with $x \equiv p/T$. Since $T(t) = T_0/a$, the left-hand side of the Boltzmann equation (11.50) simply becomes $\partial f(x,t)/\partial t$. Further, we can introduce a new time variable through

$$dy = \frac{n_eT_e\sigma_{\rm T}}{m_e}dt. \tag{11.55}$$

With this, the Boltzmann equation becomes particularly simple:

$$\frac{\partial}{\partial y} f(x, y) = \frac{1}{x^2} \frac{\partial}{\partial x} \left[ x^4 \frac{\partial}{\partial x} f(x, y) \right]. \tag{11.56}$$

Now consider the setup in Fig. 11.10, which shows a cloud of ionized gas. Before encountering the gas, the CMB photons follow an equilibrium distribution with temperature $T = T_0/a$ (we neglect the small primordial anisotropies here), so that our initial condition for the Boltzmann equation is

$$f(x, y = 0) = f^{(0)}(p = xT(t), t) = \frac{1}{e^x - 1}. \tag{11.57}$$

The photon distribution after passing through the cloud of gas can be calculated by solving Eq. (11.56). Let us consider the most relevant case in practice, when $y \ll 1$, i.e. the regime of small optical depth for Compton scattering within the cloud. The result is

$$f(x, y) \overset{y \ll 1}{=} \left\{ 1 + y \frac{1}{x^2} \frac{\partial}{\partial x} \left[ x^4 \frac{\partial}{\partial x} \right] \right\} \frac{1}{e^x - 1}, \tag{11.58}$$

where, using that $a d\chi = dt$ for photons,

$$y = \frac{\sigma_{\mathrm{T}}}{m_e} \int a n_e T_e d\chi. \tag{11.59}$$

That is, $y$ is the integral over the kinetic energy density, or pressure, in the ionized gas along the line of sight. The distortion induced by scattering captured by the second term in Eq. (11.58) differs from the distortion due to a chemical potential. Indeed, this is the first time we encounter a non-black-body photon distribution in this book, and it appears here because of the large difference between the gas temperature $T_e$ and the photon temperature $T$. This "$y$-type spectral distortion" is illustrated in Fig. 11.11. After having passed through the ionized cloud, the photon distribution has more photons in the high-energy (Wien) tail than the equilibrium distribution, and fewer of them in the low-energy (Rayleigh–Jeans) regime. We already expected this: the CMB photons are upscattered by high-energy electrons, so they are moved from the low- to the high-energy side of the distribution.

What happens when $y$ becomes very large? If we had taken a slightly more careful, longer route to derive Eq. (11.54) (not making the approximation that the photon momentum is negligible compared to that of electrons), we would have found that Eq. (11.54) involves $T_e - T$ instead of $T_e$. In the limit of many scatterings (large $y$), then, the photon distribution approaches an equilibrium distribution with a temperature dictated by energy conservation: initially, the energy density was $\rho_\gamma(T_0/a) + \rho_e(T_e)$; this must be equal to $\rho_\gamma(T_f) + \rho_e(T_f)$, thereby setting the final temperature $T_f$. Thus, our result Eq. (11.58) corresponds to the first step in the equilibration (also called "Comptonization") to the new temperature $T_f$.

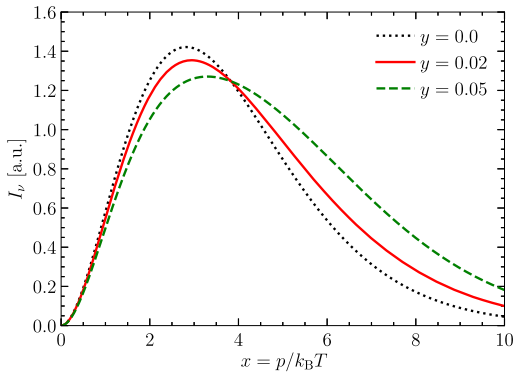



**FIGURE 11.11** Observed spectrum of CMB radiation intensity $I_\nu \propto \nu^3 f(\nu)$ with and without ($y = 0$) spectral distortion due to the SZ effect. When $y > 0$, the observed CMB has more photons in the high-energy tail and less in the low-energy tail compared to the undistorted black-body spectrum.

Eqs. (11.58)–(11.59) show that, by observing the spectral distortion in the CMB photons that pass through ionized gas, we can measure the integral of the pressure along the line of sight. Now imagine making a map of the spectral distortion across the sky. This map of $y(\hat{\boldsymbol{n}})$ corresponds to a map of the integrated pressure in ionized gas in the universe since reionization. Let us write the gas pressure in the universe as a time-dependent background with perturbations,

$$\mathcal{P}_{\rm gas}(\boldsymbol{x}, \eta) = \overline{n_e T_e}(\eta)[1 + \delta_{\mathcal{P}}(\boldsymbol{x}, \eta)]. \tag{11.60}$$

Then, the observed $y$ (Eq. (11.59)) is an integral over $\mathcal{P}_{\rm gas}$ which is analogous to the projected galaxy density field in Eq. (11.39):

$$y(\hat{\boldsymbol{n}}) = \frac{\sigma_{\rm T}}{m_e} \int_0^{\chi_*} d\chi\, \overline{n_e T_e}\, a \left[1 + \delta_{\mathcal{P}}\left(\boldsymbol{x} = \hat{\boldsymbol{n}}\chi, \eta = \eta_0 - \chi\right)\right]. \tag{11.61}$$

We can then use the results of the previous section leading to the angular power spectrum Eq. (11.47) to obtain the angular power spectrum of the SZ $y$-distortion of the CMB:

$$C_y(l) = \left(\frac{\sigma_{\rm T}}{m_e}\right)^2 \int \frac{d\chi}{\chi^2} \left[\overline{n_e T_e} a\right]^2 P_{\mathcal{P}}\left(k = \frac{l + 1/2}{\chi}, \eta(\chi)\right). \tag{11.62}$$

where $P_{\mathcal{P}}$ is the power spectrum of fractional pressure perturbations in the ionized gas. Measuring the anisotropies of the spectral distortions thus allows us to measure the amplitude of pressure fluctuations of the ionized gas in the universe, a quantity that allows for valuable insights into the thermal state of baryons which is difficult to predict theoretically. Another, more prominent, use of the SZ effect is to search for outliers of very large $y$. These correspond to rare, massive galaxy clusters (which are the topic of Sect. 12.5). One major advantage of the SZ effect over other ways to look for galaxy clusters is that the signal only

decreases weakly with distance, thanks to the use of the ubiquitous CMB as a backlight.[3] Large experimental efforts are under way that are devoted to measuring the SZ effect.

In this section, we discussed the upscattering of CMB photons due to the thermal velocities of hot electrons ("thermal SZ"). A similar effect exists due to the bulk motion of gas, which is closer to the effect of gas velocities on CMB photons discussed in Ch. 5. This effect, known as "kinetic SZ," probes the bulk momentum of gas along the line of sight rather than the pressure. It is much more difficult to detect than the thermal SZ, both due to its smaller amplitude and the fact that, since the velocity along the line of sight can have either sign, it typically cancels out unless carefully extracted using independent estimates of the gas velocity.

## 11.4 Summary

The clustering of galaxies is one of the main probes of the large-scale structure in the universe. While we spoke of galaxies throughout to be specific, a variety of tracers can be used for this purpose, and our results apply generally to any tracer. Surveys that map the distribution of tracers can be either photometric, yielding many objects with uncertain distances, or spectroscopic, with fewer objects that are, however, precisely localized. Spectroscopic surveys allow us to measure the three-dimensional statistics, in particular the power spectrum, of tracers. These are a rich source of information. However, care needs to be taken when interpreting them:

***Redshift-space distortions*** due to the peculiar velocities of galaxies lead to a characteristic dependence of the power spectrum on the angle $\mu$ of the wavevector with the line of sight. These distortions allow us to measure the amplitude of velocities, and with that, the rate of structure formation $f\sigma_8$, where $f = d\ln D_+/d\ln a$.

When analyzing galaxy surveys, we need to assume a distance-redshift relation in order to turn observed positions on the sky and redshifts into three-dimensional locations. An incorrect assumed distance-redshift relation leads to ***Alcock–Paczyński distortions***, which again are anisotropic but allow us to infer the distance to a given redshift, as well as the Hubble rate at that redshift. This would be quite difficult if the power spectrum was perfectly smooth. However, the ***BAO feature*** imprinted into matter after recombination provides a standard ruler which allows for a very precise measurement of the distance-redshift relation. Unlike standard candles such as supernovae, the BAO feature does not have to be calibrated using a distance ladder measured in the nearby universe.

Photometric galaxy data sets still allow us to measure the ***projected clustering***, the angular correlation function $w_g(\theta)$ and its Fourier counterpart, the multipoles $C_g(l)$. While the BAO feature and RSD are smoothed out, the shape and amplitude still contain cosmological information, especially when combined with weak lensing (Ch. 13).

The results in this section were based on a very important assumption about galaxy clustering: we posited that the galaxy power spectrum (in the absence of the above-

[3] Although, from Fig. 11.10, the term "backlight" is perhaps misleading: it is not only CMB photons emitted from behind the cluster that get scattered toward us, but from all directions.

mentioned distortions) is simply proportional to that of matter on large scales, with an additive noise term. This allowed us to continue to work using linear perturbation theory, resulting in the final expression for the observed galaxy power spectrum:

$$P_{\rm g,obs}(\boldsymbol{k}_{\rm obs},\bar{z}) = P_{\rm L}(k,\bar{z})\left[b_1 + f\mu_k^2\right]^2\Big|_{\boldsymbol{k}=([1+\alpha_\perp]k^1_{\rm obs},\,[1+\alpha_\perp]k^2_{\rm obs},\,[1+\alpha_\parallel]k^3_{\rm obs})} + P_N\,. \tag{11.63}$$

We did not justify why $b_1$ and $P_N$ are all we need in linear theory, which seems very simplistic given the real-world complexities of galaxy formation; yet, we will see that it holds in the next chapter, when we think about the effects of nonlinear structure formation.

Finally, we studied another important probe of structure, by going back to the Boltzmann equation for photons of Ch. 5: the spectral distortions of the CMB induced by scattering off hot ionized gas, known as ***Sunyaev–Zel'dovich (SZ) effect***. This effect can be distinguished from primordial CMB anisotropies because it modifies the CMB spectrum from its black-body form. We saw that the amplitude of observed distortions is directly proportional to the integral over the pressure in the ionized gas, allowing us to create an integrated pressure map and to identify distant massive galaxy clusters. The science we can do with clusters is one of the topics of the next chapter.

## Exercises

**11.1** Suppose the correlation function is defined as

$$\xi(\boldsymbol{r}) \equiv \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle. \tag{11.64}$$

By Fourier expanding each of the $\delta$ and using Eq. (C.22), show that this definition implies that the correlation function is the Fourier transform of the power spectrum.

**11.2** Compute (numerically) the linear growth rate $f$ today in a $\Lambda$CDM universe and compare with the approximation $f(z=0) = \Omega_{\rm m}^{0.55}$. What is the fractional error between the approximation and the exact result? Do the same for a dark energy model with $w = -0.5$.

**11.3** Compute the RMS line-of-sight velocity $\sqrt{\langle u_\parallel^2\rangle}$, in the same way as Exercise 8.13 computes the variance of the density field, i.e. using a real-space tophat filter. Use Eq. (11.17). Plot the result as a function of the filter radius $R$ at $z = 0, 0.5, 1$. Now compute the corresponding RMS value of the displacement into redshift space, which we showed to be $u_\parallel/aH$. Using that result, at what wavenumber do you roughly expect the transition between the two regimes shown in Fig. 11.3?

**11.4** The redshift-space power spectrum is often expanded in multipole moments defined through

$$P^{(l)}_{\rm g,obs}(k) = \frac{2l+1}{2}\int_{-1}^{1} d\mu_k \mathcal{P}_l(\mu_k) P_{\rm g,obs}(k,\mu_k) \tag{11.65}$$

where $P^{(l)}_{\mathrm{g,obs}}(k)$ is the $l$th multipole moment of the power spectrum. Using the orthogonality of the Legendre polynomials (Eq. (C.2)), show that this implies

$$P_{\mathrm{g,obs}}(k,\mu_k) = \sum_l \mathcal{P}_l(\mu_k) P^{(l)}_{\mathrm{g,obs}}(k). \tag{11.66}$$

Show further how the monopole and quadrupole are related to the linear matter power spectrum. In all cases, restrict to RSD and neglect the AP distortion parameters $\alpha_\parallel, \alpha_\perp$.

**11.5** In the text we showed how redshift-space distortions affect the power spectrum. Show how the redshift-space distortions affect the correlation function in the flat-sky approximation.

**11.6** Prove Eq. (11.46). Begin by rewriting the three-dimensional Dirac delta in spherical polar coordinates, and using the completeness of the spherical harmonics:

$$\begin{aligned}\delta^{(3)}_{\mathrm{D}}(\boldsymbol{x}-\boldsymbol{x}') &= \frac{1}{x^2}\delta^{(1)}_{\mathrm{D}}(x-x')\delta^{(S^2)}_{\mathrm{D}}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}') \\ &= \frac{1}{x^2}\delta^{(1)}_{\mathrm{D}}(x-x')\sum_{lm} Y_{lm}(\hat{\boldsymbol{x}})Y^*_{lm}(\hat{\boldsymbol{x}}').\end{aligned} \tag{11.67}$$

The second line follows from the fact that integrating any function defined on the unit sphere against $\delta^{(S^2)}_{\mathrm{D}}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}')$ should yield its value at $\hat{\boldsymbol{x}}$, which in turn can be expressed in terms of a sum over $Y_{lm}(\hat{\boldsymbol{x}})$ with the corresponding multipole coefficient of the function. Next, use the Fourier-space expression of the Dirac delta,

$$\delta^{(3)}_{\mathrm{D}}(\boldsymbol{x}-\boldsymbol{x}') = \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')}, \tag{11.68}$$

along with the spherical decomposition of the exponential, Eq. (C.17) (twice). Finally, argue that the fact that Eq. (11.67) has to hold in general implies that Eq. (11.46) holds individually for all $l = 0, 1, \ldots$.

**11.7** **(a)** Plot the integrand of Eq. (11.45) as a function of $k$ and $\chi$ assuming $\chi = \chi'$ for different $l$. What is the approximate scaling with $l$ of the peak position and width for $l \gg 1$? What happens if $\chi \neq \chi'$?

**(b)** Evaluate the exact integral Eq. (11.44) and the Limber approximation Eq. (11.47) for $l = 2, 5, 10, 30$. Assume that $W(\chi)$ is a Gaussian centered around $\chi(z=1)$ with RMS corresponding to a redshift uncertainty of $\Delta z = 0.2$. Determine empirically the accuracy of the Limber approximation as a function of $l$.

**11.8** Decompose the galaxy angular correlation function into a sum over multipole moments,

$$w_{\mathrm{g}}(\theta) = \sum_{l=1}^{\infty} \frac{2l+1}{4\pi} C_{\mathrm{g}}(l)\mathcal{P}_l(\cos\theta). \tag{11.69}$$

Express $C_g(l)$ as an integral over the 3D galaxy power spectrum $P_{g,obs}(\boldsymbol{k})$. Show that on small scales, the $C_g(l)$ defined in this way are equal to those derived in Sect. 11.2.

**11.9** Derive the steps leading from Eq. (11.51) to Eq. (11.54).

**11.10** **(a)** Show that the integral over all $p$ of the collision term in Eq. (11.54) vanishes, so that the photon number is conserved.

**(b)** Evaluate the right-hand side of Eq. (11.58) explicitly and show that it corresponds to a spectral distortion, i.e. cannot be absorbed by a modified temperature and chemical potential.

**11.11** The SZ effect enhances the observed spectrum at high frequencies, and suppresses it at low frequencies. Find the frequency at which the distortion vanishes ("SZ null"). Show that setting the derivative in the second term of Eq. (11.58) equal to zero leads to

$$(4-x)e^x = 4+x. \tag{11.70}$$

You can solve this numerically or note that the left-hand side reaches a maximum at $x=3$, where it is much larger than the right-hand side, and then drops to zero at $x=4$. So, the place where the two sides are equal is when $x$ is a bit below 4. Expand perturbatively around $x=4$ to obtain an estimate of this solution. Compare with the actual solution of $x=3.83$. What frequency does that correspond to for the CMB with $T=2.726$ K?

# 12

# Growth of structure: beyond linear theory

So far in this book, we have focused on small perturbations to a homogeneous universe; technically, we have worked to linear order in all perturbations, such as those in radiation and matter, as well as spacetime perturbations. That was sufficient for an accurate description of the CMB, but clearly fails in describing the late universe, with its stars, galaxies, galaxy clusters, and so on.

We thus want to extend our model of the universe beyond linear order. However, the fundamental equations we introduced in Ch. 3—the Einstein and Boltzmann equations—are extremely complex in full generality. Fortunately, not all is lost. Even in the late universe with nonlinear structures, gravity is still weak in a certain sense: the metric remains close to FLRW almost everywhere. This allows us to continue to work to linear order in metric perturbations, thus greatly reducing the complexity of Einstein's equations, while being fully nonlinear in the matter density.

The clustering components of the universe are dominated by dark matter. Moreover, pressure forces of the gas (baryons) become relevant only on very small scales, due to the relative coldness of the gas, which cools rapidly after recombination. Thus, approximating baryons as collisionless, we will focus on solving the evolution of cold, collisionless matter under gravity in this chapter. This is in fact a beautiful, conceptually simple problem. We will introduce two main tools to solve it: perturbation theory and simulations. A qualitative prediction can, however, already be deduced from the shape of the linear matter power spectrum (Fig. 8.3, and, in a more suitable representation, Fig. 12.1): density perturbations on large scales are small, while those on small scales are large. Hence, structure becomes nonlinear (collapses) first on small scales. These small-scale collapsed structures then progressively ("hierarchically") assemble to larger structures as the universe evolves. The nonlinear structure in the universe can be thought of as being made up of bound dark matter structures referred to as *halos*, which form useful building blocks for an empirical understanding of the complex nonlinear matter distribution.

Next, we turn to galaxies. Unlike collisionless matter, the formation of galaxies is not "simply" described by the collisionless Boltzmann and Einstein equations. Instead, they form through the radiative and collisional cooling of gas, which eventually collapses to form stars. Despite these complexities, we will see that perturbative approaches can still be used to describe galaxy clustering on large scales. This is crucial in order to be able to use galaxy clustering as a cosmological probe using the techniques we studied in the preceding chapter: in particular the BAO standard ruler, AP and redshift-space distortions.

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00018-8
Copyright © 2021 Elsevier Inc. All rights reserved.

Finally, another way of gleaning information about the underlying mass density is by studying clusters of galaxies. Clusters can now be probed with many different techniques, as we will see in Sect. 12.5. Further, they are fairly faithful tracers of the most massive dark matter halos, whose abundance is a sensitive probe of structure formation. Therefore, counting clusters leads to interesting cosmological constraints as well.

This is a very substantial chapter, which, after an overview in Sect. 12.1, covers a range of topics that at first sight might appear disparate. So a brief guide to this chapter is in order. The two fundamental approaches—perturbation theory and numerical simulations—are covered in Sect. 12.2 and Sect. 12.3, respectively. If you are interested in understanding the assumptions made in Ch. 11 about the clustering of galaxies on large scales, you should work through Sect. 12.2 and Sect. 12.6. These are independent of the section on simulations. If you are mostly interested in galaxy clusters, you can focus on Sect. 12.4 which covers dark matter halos, in addition to the cluster section 12.5. Finally, section 12.7 on the semi-analytic halo model relies only on Sect. 12.4 and the clustering assumptions made in Ch. 11.

The topics covered in this chapter are not prerequisites for the following chapters, although we will see that a solid theoretical prediction for the nonlinear matter distribution is a key requirement in order to infer cosmology from gravitational lensing (Ch. 13).

## 12.1 Prelude

The dominant clustering component in the late universe is matter, which consists of dark matter (to about 80%) as well as baryons (in the form of neutral and ionized gas as well as stars). In the bulk of this chapter, we will lump together dark matter and baryons, and refer to them simply as "matter." Of course, baryons behave differently from dark matter as they feel electromagnetic forces. However, after having completely decoupled from the photons, baryons cool rapidly (their temperature scales as kinetic energy, so $\propto a^{-2}$), so that the pressure induced by electromagnetic interactions is actually only relevant on very small scales. For this reason, a good and practical approximation is to consider all of matter as a single component while neglecting all non-gravitational forces. This means that we will start from the equations for dark matter, but now also include baryons.[1]

Let us go back to the equations for the linear evolution of dark matter derived in Sect. 5.4 and Sect. 6.3.2. These consist of the continuity, Euler, and Poisson equations:

$$\begin{aligned} {\delta_{\rm m}}' + iku_{\rm m} + 3\Phi' &= 0, \\ {u_{\rm m}}' + \frac{a'}{a}u_{\rm m} + ik\Psi &= 0, \\ k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) &= 4\pi G a^2 \rho_{\rm m}\delta_{\rm m}, \end{aligned} \tag{12.1}$$

[1] We will neglect the different initial conditions for dark matter and baryons we studied in Sect. 8.6.1 in this chapter. They can be similarly treated using the techniques we describe here, and only lead to percent-level corrections at late times.

where on the right-hand side of the Poisson equation we have only included matter. This is justified since at redshifts $z \lesssim 10$, where structure begins to become nonlinear, the contribution of radiation is negligible. This is not entirely true for neutrinos, which contribute at the percent-level, because they have a finite mass and thus become non-relativistic. Including neutrinos, however, would not change our main arguments in the following. Further, we can set $\Phi = -\Psi$, since the late universe has negligible anisotropic stress.[2]

Before jumping into the calculation of nonlinear growth, let us pause to consider which scales we will be dealing with. After having solved the linear evolution equations in Ch. 8, we are able to calculate the typical amplitude of linear matter density fluctuations on a given scale. Let us define the filtered density field $\delta_W(\boldsymbol{x})$,

$$\delta_W(\boldsymbol{x}) = \int d^3y\, W(|\boldsymbol{x}-\boldsymbol{y}|)\delta_{\rm m}(\boldsymbol{y}), \tag{12.2}$$

where $W(x)$ is the filtering kernel that we can take to be isotropic so that it only depends on the magnitude of $\boldsymbol{x}-\boldsymbol{y}$. This filtering corresponds to a multiplication in Fourier space:

$$\delta_W(\boldsymbol{k}) = W(k)\delta_{\rm m}(\boldsymbol{k}), \tag{12.3}$$

where $W(k)$ is the Fourier transform of the isotropic filtering kernel (in this chapter, we will move back and forth between real and Fourier space. Any ambiguity is removed, however, by the arguments of functions or the explicit appearance of factors of $\boldsymbol{k}$). Notice that a filter that is normalized in real space via $\int d^3x\, W(\boldsymbol{x}) = 1$ obeys $W(k=0) = 1$ in Fourier space. It could be, for example, a Gaussian with width $\Delta k$, which corresponds to a Gaussian with width $R = 1/\Delta k$ in real space. Then, the variance of this filtered density field is directly related to the matter power spectrum (as you can derive in Exercise 8.13):

$$\begin{aligned}\sigma_W^2 \equiv \left\langle (\delta_W)^2(\boldsymbol{x})\right\rangle &= \int \frac{d^3k}{(2\pi)^3}\int \frac{d^3k'}{(2\pi)^3}\left\langle \delta_W(\boldsymbol{k})\delta_W^*(\boldsymbol{k}')\right\rangle e^{i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}} \\ &= \int \frac{d^3k}{(2\pi)^3} P_{\rm L}(k)|W(k)|^2 \\ &= \frac{1}{2\pi^2}\int d\ln k\, k^3 P_{\rm L}(k)|W(k)|^2. \end{aligned} \tag{12.4}$$

The result is shown in Fig. 12.1: when smoothed on a large scale, density fluctuations are small, while they become large when we filter on a smaller scale. For a sufficiently small filter scale, $\sigma_W^2$ becomes greater than 1. This means that, when we look at our universe on a sufficiently small scale, i.e. with sufficiently high resolution, any given point is likely to have a density that is very different from the cosmic mean. That means that our linear treatment based on Eq. (12.1) predicts a wrong result for the density field in most places.

[2] We choose to work with $\Psi$ in the following, as the perturbation to the time-time component of the metric is what physically governs the motion of non-relativistic matter. When comparing to the literature, keep in mind that different notation (e.g., $\Phi$ instead of $\Psi$) and different sign conventions are common.

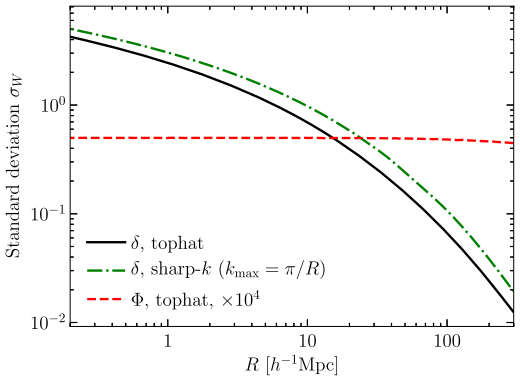



**FIGURE 12.1** Standard deviation $\sigma_W = \sqrt{\langle \delta_W^2 \rangle}$ of the linear matter density field at $z = 0$ smoothed with real-space tophat and sharp-$k$ filters, as a function of the smoothing scale $R$. When filtered on a large scale, the fluctuations of the density field are small, while fluctuations on small scales become large. Evaluating the black solid curve at $R = 8\,h^{-1}\,\mathrm{Mpc}$ yields the commonly used amplitude parameter $\sigma_8$. We also show the RMS value $\sqrt{\langle \Phi_W^2 \rangle}$ of the gravitational potential multiplied by $10^4$. The potential fluctuations are very small on *all* scales.

Clearly, we have to do better. Notice also that the precise filter shape is not important for this conclusion, as any reasonable filter leads to the same trend.

We can also compute the variance of metric perturbations $\Psi$ as a function of scale. This is also shown in the figure. Interestingly, the typical potential fluctuations remain small, $\lesssim 10^{-4}$, on *all scales*.[3] This is easy to understand: the integral in Eq. (12.4) is dominated by high wavenumbers $k$, and peaks near the scale picked out by the filter $W$. On small scales, then, the integral is dominated by contributions where $k \gg aH \sim 3 \cdot 10^{-4}\,h\ \mathrm{Mpc}^{-1}$, that is, spatial scales that are much smaller than the Hubble radius. Then, the first term in the Poisson equation (12.1) is by far the dominant one (note that $\Phi'$ is at most of order $(a'/a)\Phi$), and it simply becomes

$$-k^2\Psi = 4\pi G a^2 \rho_{\mathrm{m}} \delta_{\mathrm{m}}. \tag{12.5}$$

This is the well-known Poisson equation of Newtonian gravity, with additional factors of $a$ because the wavenumber $k$ is in comoving units. Thus, the magnitude of $\Psi(\boldsymbol{k})$ is proportional to $\delta_{\mathrm{m}}(\boldsymbol{k})/k^2$, and so is highly suppressed compared to the density on small scales. This explains why the typical potential fluctuations in the universe remain small even though density fluctuations become large. Another way to see the same result is to recall the evolution of potential and density during matter domination: the potentials remain constant, while the density perturbations grow as the linear growth factor $D_+(\eta) \propto a(\eta)$.

We can use this result to our advantage. First, given the smallness of spacetime perturbations, we can continue to work to linear order in the potential $\Psi$. This means that the

[3] Technically, $\left\langle \Phi_W^2 \right\rangle$ diverges logarithmically when including modes with $k \to 0$. Only potential perturbations within our current horizon are observable, so we have used a cutoff $k_{\min} = 10^{-4}\,h\,\mathrm{Mpc}^{-1}$. The precise value of this cutoff has a very small impact on the numerical result.

linear-order Einstein equations we have derived in Ch. 6 are sufficient. Second, since nonlinear evolution is relevant only on small scales (compared to the Hubble radius), we can employ the Newtonian limit of the relevant Einstein equation, i.e. Eq. (12.5). This greatly simplifies the gravity side of the problem, and we can devote our attention to the dynamics of matter. The latter approximation is better than it seems: Eq. (12.5) retains its validity on *all scales* in matter domination if $\delta_{\rm m}$ is the density perturbation in *synchronous-comoving gauge*. The latter coordinates are defined by $g_{00} = -1$ (i.e. no time-time perturbation, so that the time coordinate is the proper time: synchronous; see Exercise 5.1), and no velocities $\boldsymbol{u}_{\rm m} = 0$ (comoving). As long as we keep this interpretation of $\delta_{\rm m}$ in mind in the following, the results of perturbation theory and simulations that we will obtain are valid on all scales, including those comparable to the horizon.

We now want to extend Eq. (12.1) to nonlinear order. To do this, let us go back to the starting point of these equations, which we obtained by taking moments of the Boltzmann equation. So we need an expression for the Boltzmann equation that is not restricted to small perturbations, but applies to non-relativistic matter on sub-horizon scales. We begin with the general collisionless Boltzmann equation written in Cartesian form:

$$\frac{df_{\rm m}}{dt} = \frac{\partial f_{\rm m}}{\partial t} + \frac{\partial f_{\rm m}}{\partial x^i}\frac{dx^i}{dt} + \frac{\partial f_{\rm m}}{\partial p^i}\frac{dp^i}{dt} = 0, \tag{12.6}$$

where $f_{\rm m}$ is the distribution function for matter. Now, using the fact that matter is moving slowly, we expand $E(p) = m + p^2/2m$ and keep only the leading terms in $p/m$. This yields $dx^i/dt = p^i/am$ from the geodesic equation. The term $dp^i/dt$ is also straightforward, starting from Eq. (3.69):

$$\begin{aligned}\frac{dp^i}{dt} &= -\left(H + \dot{\Phi}\right)p^i - \frac{E}{a}\Psi_{,i} - \frac{1}{a}\frac{p^i}{E}p^k\Phi_{,k} + \frac{p^2}{aE}\Phi_{,i} \\ &\to -Hp^i - \frac{m}{a}\Psi_{,i} \qquad \text{(non-relativistic, sub-horizon)}.\end{aligned} \tag{12.7}$$

All other terms are either suppressed on small scales ($\dot{\Phi}$) or negligible due to the small velocities (terms of order $p^2/E$). Inserting these results into the Boltzmann equation, we obtain

$$\frac{df_{\rm m}}{dt} = \frac{\partial f_{\rm m}}{\partial t} + \frac{\partial f_{\rm m}}{\partial x^j}\frac{p^j}{ma} - \frac{\partial f_{\rm m}}{\partial p^j}\left[Hp^j + \frac{m}{a}\frac{\partial\Psi}{\partial x^j}\right] = 0. \tag{12.8}$$

Let us recap the significance of this result: Eq. (12.8) *does not* assume that the distribution function is close to its value in the homogeneous universe. It *does* assume small spacetime perturbations, which we have found to be an excellent approximation on all scales. The same reasoning that simplified the 00-component of the Einstein equation to Eq. (12.5) allowed us to drop the $\dot{\Phi}$ term, which is at most of order $aH\Psi$, and hence much smaller than the $\partial\Psi/\partial x^j$ contribution, which is of order $k\Psi$.

The coupled set of Eq. (12.8) and Eq. (12.5) forms the starting point for the nonlinear evolution of matter. It is known as the *Vlasov–Poisson* system. A nonlinear system (through

the coupling between $\Psi$ and $f_\mathrm{m}$) of integro-differential equations (because $\delta_\mathrm{m}$ is an integral over the distribution function $f_\mathrm{m}$) in $6+1$ dimensions, it is notoriously difficult to solve. The following sections will deal with perturbative as well as numerical techniques to solve it.

The perturbative approach proceeds as we have done in previous chapters: by taking moments of the Boltzmann equation. In the linear regime that we studied so far, the distribution function $f_\mathrm{m}$ was completely described by its zeroth (density) and first moments (velocity). Physically, this means that the second moment, the velocity dispersion, is vanishingly small. Then the distribution function can be written as

$$f_\mathrm{m}(\boldsymbol{x},\boldsymbol{p},t) = \frac{\rho_\mathrm{m}(\boldsymbol{x},t)}{m}(2\pi)^3\delta_\mathrm{D}^{(3)}(\boldsymbol{p}-m\boldsymbol{u}_\mathrm{m}(\boldsymbol{x},t)) \qquad \text{(no velocity dispersion)}, \tag{12.9}$$

where we have absorbed the irrelevant degeneracy factors of CDM and baryon species into $f_\mathrm{m}$. You can think of this as arising from a thermal velocity distribution at each point centered around $\boldsymbol{u}_\mathrm{m}(\boldsymbol{x},t)$ when taking the limit of zero temperature. It is important to realize, however, that the form of the distribution function Eq. (12.9) does not remain valid once structure becomes nonlinear. We will study in more detail how this happens in Sect. 12.3. First though, let us see how far we get with the ansatz of vanishing velocity dispersion.

## 12.2 Perturbation theory

The starting point of perturbative approaches to the nonlinear growth of structure is to take moments of the Vlasov equation; that is, we follow the same basic approach we took in Ch. 5. For any function $A(\boldsymbol{x},\boldsymbol{p},t)$ defined on $6+1$ dimensional phase space, we can define the momentum average

$$\langle A\rangle_{f_\mathrm{m}}(\boldsymbol{x},t) \equiv \int \frac{d^3p}{(2\pi)^3}A(\boldsymbol{x},\boldsymbol{p},t)\,f_\mathrm{m}(\boldsymbol{x},\boldsymbol{p},t), \tag{12.10}$$

which now is only a function of position and time. Again, we absorb any degeneracy factors into $f_\mathrm{m}$; there are no collision terms where they could become relevant. Choosing $A=1$ then simply gives us the number density:

$$\langle 1\rangle_{f_\mathrm{m}}(\boldsymbol{x},t) = n(\boldsymbol{x},t) = \frac{\rho_\mathrm{m}(\boldsymbol{x},t)}{m}. \tag{12.11}$$

Equivalently, $\langle m\rangle_{f_\mathrm{m}}$ yields the mass density $\rho_\mathrm{m}(\boldsymbol{x},t)$, which is more useful in practice. Similarly, we define the bulk or fluid velocity as the momentum average of $p^i$, normalized by the density:

$$u^i_\mathrm{m}(\boldsymbol{x},t) \equiv \frac{\langle p^i\rangle_{f_\mathrm{m}}}{\langle m\rangle_{f_\mathrm{m}}}. \tag{12.12}$$

Let us now take the momentum average $\int d^3p/(2\pi)^3$ of the Vlasov equation (12.8), multiplied by $m$, thus taking the zeroth moment of the Vlasov equation. We can always pull out

derivatives with respect to $t$ and $\boldsymbol{x}$ outside the momentum integral, to obtain

$$\frac{\partial}{\partial t}\rho_{\rm m}(\boldsymbol{x},t)+\frac{1}{a}\frac{\partial}{\partial x^j}\left[\rho_{\rm m}(\boldsymbol{x},t)u^j_{\rm m}(\boldsymbol{x},t)\right]-\int\frac{d^3p}{(2\pi)^3}m\left[Hp^j+\frac{m}{a}\frac{\partial\Psi}{\partial x^j}\right]\frac{\partial}{\partial p^j}f_{\rm m}(\boldsymbol{x},\boldsymbol{p},t)=0, \tag{12.13}$$

where we have used that $\left\langle p^j\right\rangle_{f_{\rm m}}=\rho_{\rm m}u^j_{\rm m}$. The last term can be integrated by parts to move the derivative with respect to $p^j$ from $f_{\rm m}$ to the term in square brackets (the boundary term vanishes, since any well-behaved distribution function does not have particles at infinite momentum). Evaluating this derivative, we obtain, first, $-\partial/\partial p^j(Hp^j)=-3H$, while $\partial/\partial p^j(\partial\Psi/\partial x^j)=0$, since the potential $\Psi$ is only a function of $t$ and $\boldsymbol{x}$. Thus, Eq. (12.13) becomes

$$\frac{\partial}{\partial t}\rho_{\rm m}(\boldsymbol{x},t)+\frac{1}{a}\frac{\partial}{\partial x^j}\left[\rho_{\rm m}(\boldsymbol{x},t)u^j_{\rm m}(\boldsymbol{x},t)\right]+3H\rho_{\rm m}(\boldsymbol{x},t)=0. \tag{12.14}$$

Modulo an overall factor $m$, this is the continuity equation whose linear version is Eq. (5.41), but now valid at fully nonlinear order (and on sub-horizon scales).

As in the linear case, Eq. (12.14) is not sufficient, since we need an equation for the velocity $u^i_{\rm m}$ as well. Let us thus take the first moment of the Vlasov equation (12.8), by multiplying with $p^i$ and integrating over $\boldsymbol{p}$:

$$\frac{\partial}{\partial t}\left[\rho_{\rm m}u^i_{\rm m}(\boldsymbol{x},t)\right]+\frac{1}{ma}\frac{\partial}{\partial x^j}\left\langle p^ip^j\right\rangle_{f_{\rm m}}-\int\frac{d^3p}{(2\pi)^3}p^i\left[Hp^j+\frac{m}{a}\frac{\partial\Psi}{\partial x^j}\right]\frac{\partial}{\partial p^j}f_{\rm m}(\boldsymbol{x},\boldsymbol{p},t)=0. \tag{12.15}$$

The last term can again be dealt with by integration by parts, and we obtain

$$\frac{\partial}{\partial t}\left[\rho_{\rm m}u^i_{\rm m}(\boldsymbol{x},t)\right]+\frac{1}{ma}\frac{\partial}{\partial x^j}\left\langle p^ip^j\right\rangle_{f_{\rm m}}+4H\rho_{\rm m}u^i_{\rm m}(\boldsymbol{x},t)+\frac{1}{a}\rho_{\rm m}(\boldsymbol{x},t)\frac{\partial\Psi(\boldsymbol{x},t)}{\partial x^i}=0. \tag{12.16}$$

This is our desired equation for $u^i_{\rm m}$, but we now encounter another quantity, the second moment of the distribution $\left\langle p^ip^j\right\rangle_{f_{\rm m}}$. Let us write this as follows, introducing the *stress tensor* $\sigma^{ij}_{\rm m}(\boldsymbol{x},t)$:

$$\frac{1}{m}\left\langle p^ip^j\right\rangle_{f_{\rm m}}=\rho_{\rm m}u^i_{\rm m}u^j_{\rm m}+\sigma^{ij}_{\rm m}. \tag{12.17}$$

As with $u^i_{\rm m}$ and $p^i$, we do not need to distinguish between upper and lower latin indices on $\sigma^{ij}_{\rm m}$. At this point, this is nothing but a definition for $\sigma^{ij}_{\rm m}$, but we will learn the significance of this decomposition in a moment. Inserting this into Eq. (12.16), we obtain

$$\frac{\partial}{\partial t}\left[\rho_{\rm m}u^i_{\rm m}(\boldsymbol{x},t)\right]+\frac{1}{a}\frac{\partial}{\partial x^j}\left[\rho_{\rm m}u^i_{\rm m}u^j_{\rm m}+\sigma^{ij}_{\rm m}\right]+4H\rho_{\rm m}u^i_{\rm m}(\boldsymbol{x},t)+\frac{1}{a}\rho_{\rm m}(\boldsymbol{x},t)\frac{\partial\Psi(\boldsymbol{x},t)}{\partial x^i}=0. \tag{12.18}$$

This equation becomes much more familiar if we subtract the continuity equation (12.14), multiplied by $u^i_{\rm m}$, from it:

$$\rho_{\rm m}\frac{\partial}{\partial t}u^i_{\rm m}+\frac{1}{a}\rho_{\rm m}u^j_{\rm m}\frac{\partial}{\partial x^j}u^i_{\rm m}+H\rho_{\rm m}u^i_{\rm m}+\frac{1}{a}\rho_{\rm m}\frac{\partial\Psi}{\partial x^i}+\frac{1}{a}\frac{\partial}{\partial x^j}\sigma^{ij}_{\rm m}=0. \tag{12.19}$$

Finally, simply dividing by $\rho_{\rm m}$ leaves us with the Euler equation in the expanding universe:

$$\frac{\partial}{\partial t}u^i_{\rm m}+\frac{1}{a}u^j_{\rm m}\frac{\partial}{\partial x^j}u^i_{\rm m}+Hu^i_{\rm m}+\frac{1}{a}\frac{\partial\Psi}{\partial x^i}+\frac{1}{\rho_{\rm m}a}\frac{\partial}{\partial x^j}\sigma^{ij}_{\rm m}=0. \tag{12.20}$$

Again, this generalizes our previous result Eq. (5.50) to nonlinear order, but restricting to sub-horizon scales. The first two terms correspond to the convective or material derivative $\partial/\partial t+u^j_{\rm m}\partial/\partial x^j$ acting on the velocity $u^i_{\rm m}$; note that this includes a nonlinear term. The third term is the Hubble drag which, in the absence of any perturbations, leads to a decay in the velocity proportional to $1/a$. The fourth term is the effect of gravity, which is the same as at linear order and looks precisely like the result in Newtonian gravity, once converting from comoving coordinates $x^i$ to physical distance intervals given by $dr^i=adx^i$. All of these look familiar, or are equivalent to what we obtained in Sect. 5.4 when restricting to linear order.

Finally, we have the contribution from the stress tensor. Assume for a moment that it is diagonal, so that we can write $\sigma^{ij}_{\rm m}(\boldsymbol{x},t)=\mathcal{P}_{\rm m}(\boldsymbol{x},t)\delta_{ij}$. Then, the last term becomes $\partial\mathcal{P}_{\rm m}/\partial x^i/(\rho_{\rm m}a)$, which is precisely the contribution of pressure to the Euler equation. Thus, $\sigma^{ij}_{\rm m}$ encodes generalized pressure forces. But shouldn't cold matter have zero pressure? Indeed, it is easy to verify (Exercise 12.1) that if we insert a "cold" distribution function of the form Eq. (12.9) into Eq. (12.17), we obtain $\sigma^{ij}_{\rm m}=0$. Thus, it is a standard assumption to drop the stress tensor from the Euler equation. Then, we are left with three equations—continuity, Euler, and Poisson—for three unknowns: $\rho_{\rm m}$, $u^i_{\rm m}$, $\Psi$, which form a closed system. Let us thus proceed in solving this set of equations, and return to the question of whether we can truly neglect $\sigma^{ij}_{\rm m}$ at the end.

Our first step is to remove the homogeneous part of the continuity equation, since the spatially constant background density does not have a dynamical effect (it does not source the gravitational potential $\Psi$). For this, we use the definition of $\delta_{\rm m}$ through $\rho_{\rm m}(\boldsymbol{x},t)=\rho_{\rm m}(t)[1+\delta_{\rm m}(\boldsymbol{x},t)]$. The continuity equation in the background is $\partial\rho_{\rm m}/\partial t+3H\rho_{\rm m}=0$, so we subtract this, multiplied by a factor $1+\delta_{\rm m}$,

$$[1+\delta_{\rm m}(\boldsymbol{x},t)]\left[\frac{\partial}{\partial t}\rho_{\rm m}(t)+3H\rho_{\rm m}(t)\right]=0, \tag{12.21}$$

from Eq. (12.14), to obtain

$$\rho_{\rm m}\frac{\partial}{\partial t}[1+\delta_{\rm m}(\boldsymbol{x},t)]+\frac{\rho_{\rm m}}{a}\frac{\partial}{\partial x^j}\left[(1+\delta_{\rm m})u^j_{\rm m}(\boldsymbol{x},t)\right]=0. \tag{12.22}$$

We can now divide by $\rho_{\rm m}$ to obtain the continuity equation relating $\delta_{\rm m}$ and $u^i_{\rm m}$. Finally, let us use conformal time. Multiplying the equation by $a$, we obtain the following set of our

three equations to solve:

$$\begin{aligned} {\delta_{\rm m}}' + \frac{\partial}{\partial x^j}\left[(1+\delta_{\rm m})u^j_{\rm m}\right] &= 0, \\ {u^i_{\rm m}}' + u^j_{\rm m}\frac{\partial}{\partial x^j}u^i_{\rm m} + aHu^i_{\rm m} + \frac{\partial\Psi}{\partial x^i} &= 0, \\ \nabla^2\Psi = \frac{3}{2}\Omega_{\rm m}(\eta)(aH)^2\delta_{\rm m}. & \end{aligned} \tag{12.23}$$

In the last line, we have used the definition of the *time-dependent* density parameter $\Omega_{\rm m}(\eta)$ to replace $4\pi G\rho_{\rm m}$ with $(3/2)\Omega_{\rm m}(\eta)H^2(\eta)$. $\Omega_{\rm m}(\eta)$ is to be distinguished from our convention $\Omega_{\rm m} = \Omega_{\rm m}(\eta_0)$ up to now. We will use $\Omega_{\rm m}(\eta)$ only in this section, since it is very convenient, and revert back to the $\Omega_{\rm m} = \Omega_{\rm m}(\eta_0)$ convention after; bear in mind, however, that the use of a time-dependent $\Omega_{\rm m}$ is quite common in the literature.

We have thus reduced the $6+1$-dimensional Vlasov–Poisson system of integro-differential equations into the Euler–Poisson system of coupled partial differential equations in $3+1$ dimensions—a significant simplification! Next, let us introduce the velocity divergence $\theta_{\rm m} \equiv \partial_i u^i_{\rm m}$, and take the divergence of the Euler equation. Further, let us move those terms that are nonlinear in the variables we wish to solve for to the right-hand side:

$$\begin{aligned} {\delta_{\rm m}}' + \theta_{\rm m} &= -\delta_{\rm m}\theta_{\rm m} - u^j_{\rm m}\frac{\partial}{\partial x^j}\delta_{\rm m}, \\ {\theta_{\rm m}}' + aH\theta_{\rm m} + \nabla^2\Psi &= -u^j_{\rm m}\frac{\partial}{\partial x^j}\theta_{\rm m} - (\partial_i u^j_{\rm m})(\partial_j u^i_{\rm m}). \end{aligned} \tag{12.24}$$

Unfortunately, this is still a coupled system of nonlinear partial differential equations which in general cannot be solved in any closed form. However, we will see that the simplicity of the matter-dominated universe allows us to make progress in an approximate way.

If we set the right-hand sides of the continuity and Euler equations to zero, we recover the linear set of equations we solved to obtain the growth factor in Sect. 8.5. That is, the solution for the density was simply proportional to the initial density field, with a time-dependent proportionality constant which we called the growth factor $D_+(\eta)$:

$$\delta_{\rm m}(\boldsymbol{x},\eta) = \delta^{(1)}(\boldsymbol{x},\eta) \equiv D_+(\eta)\delta_0(\boldsymbol{x}), \tag{12.25}$$

where $\delta_0(\boldsymbol{x}) = \delta_{\rm m}(\boldsymbol{x},\eta_{\rm ref})/D_+(\eta_{\rm ref})$ is the scaled density field at some arbitrary, but fixed reference epoch $\eta_{\rm ref}$. The linear continuity equation yields

$$\theta^{(1)}(\boldsymbol{x},\eta) = -{\delta^{(1)}}'(\boldsymbol{x},\eta) = -aHf(\eta)\delta^{(1)}(\boldsymbol{x},\eta), \tag{12.26}$$

where $f = d\ln D_+/d\ln a$ is the growth rate we introduced in Sect. 8.5. You might have noticed a subtle assumption we have made in going from Eq. (12.23) to Eq. (12.24): by taking the divergence of the Euler equation, we have neglected the curl part of the velocity, or vorticity, $\boldsymbol{\omega}_{\rm m} = \boldsymbol{\nabla}\times\boldsymbol{u}_{\rm m}$. As we have seen in Ch. 8, the growing-mode solution Eq. (12.25)

corresponds to a longitudinal, i.e. curl-free velocity field. In fact, you can show in Exercise 12.2 that the vorticity is not sourced in the system Eq. (12.23) even at nonlinear order. This means that it keeps decaying ($\omega^i_{\rm m} \propto 1/a$) and can be neglected here.

Eq. (12.24) now suggests an iterative approach to the solution: our linear solution was obtained neglecting the nonlinear terms on the right-hand side. Our next approximation is to insert the linear solution into the nonlinear terms:

$$\begin{aligned} \delta^{(2)\prime} + \theta^{(2)} &= -\delta^{(1)}\theta^{(1)} - (u^{(1)})^j \frac{\partial}{\partial x^j}\delta^{(1)}, \\ \theta^{(2)\prime} + aH\theta^{(2)} + \frac{3}{2}\Omega_{\rm m}(\eta)(aH)^2\delta^{(2)} &= -(u^{(1)})^j \frac{\partial}{\partial x^j}\theta^{(1)} - [\partial_i (u^{(1)})^j][\partial_j (u^{(1)})^i], \end{aligned} \tag{12.27}$$

where we have used the Poisson equation for $\Psi^{(2)}$,

$$\nabla^2 \Psi^{(2)} = \frac{3}{2}\Omega_{\rm m}(\eta)(aH)^2\delta^{(2)}. \tag{12.28}$$

This is now an inhomogeneous but still linear system of partial differential equations for $\delta^{(2)}$, $\theta^{(2)}$. In fact, it can be turned into a system of *ordinary* differential equations and then solved. We will see how this miracle happens in a moment. Eq. (12.27) shows that $\delta^{(2)}$ and $\theta^{(2)}$ are sourced by terms that involve the square of the linear fields. Then, on large scales where these linear fields are small (see Fig. 12.1), the source terms will be even smaller so that $\delta^{(2)}$ is a small correction to $\delta^{(1)}$. The end result we are aiming for, then, is to expand the nonlinear field $\delta_{\rm m}$ as

$$\begin{aligned} \delta_{\rm m}(\boldsymbol{x}, \eta) &= \delta^{(1)}(\boldsymbol{x}, \eta) + \delta^{(2)}(\boldsymbol{x}, \eta) + \cdots + \delta^{(n)}(\boldsymbol{x}, \eta), \\ \theta_{\rm m}(\boldsymbol{x}, \eta) &= \theta^{(1)}(\boldsymbol{x}, \eta) + \theta^{(2)}(\boldsymbol{x}, \eta) + \cdots + \theta^{(n)}(\boldsymbol{x}, \eta), \end{aligned} \tag{12.29}$$

where the source terms for $\delta^{(n)}$, $\theta^{(n)}$ involve $n$ powers of the linear fields, and so each term in the series Eq. (12.29) is smaller than the previous one. As long as this holds, our perturbation-theory prediction for $\delta_{\rm m}$ and $\theta_{\rm m}$ should become more and more accurate as we increase $n$, i.e. include more higher-order terms. Computing the terms in the expansion Eq. (12.29), and determining the scales on which this expansion is valid, are the main goals of the perturbative approach to nonlinear large-scale structure. Notice that, starting from Eq. (12.25), we have dropped the subscripts "m" on $\delta^{(n)}$, $\theta^{(n)}$ for notational clarity, since we deal exclusively with the matter fields in the following.

To begin, let us transform Eq. (12.27) to Fourier space, $\boldsymbol{x} \to \boldsymbol{k}$. The left-hand sides are easy to transform, since they are linear. The real-space products on the right-hand side turn into convolutions in Fourier space, where the linear density, velocity, and potential are simply related in Fourier space:

$$\begin{aligned} (u^{(1)})^i(\boldsymbol{k}, \eta) &= \frac{ik^i}{k^2} aHf\delta^{(1)}(\boldsymbol{k}, \eta), \\ \Psi(\boldsymbol{k}, \eta) &= -\frac{3}{2}\Omega_{\rm m}(\eta)\frac{(aH)^2}{k^2}\delta_{\rm m}(\boldsymbol{k}, \eta). \end{aligned} \tag{12.30}$$

Very importantly, the last relation holds not only for $\Psi^{(1)}$, but at any order, because the Poisson equation is linear (as we are in the weak-gravity regime); we already used this fact for $\Psi^{(2)}$. We thus obtain, using Eq. (12.25),

$$\begin{aligned}
\delta^{(2)\prime}(\boldsymbol{k},\eta)+\theta^{(2)}(\boldsymbol{k},\eta) &= \int \frac{d^3k_1}{(2\pi)^3}\int \frac{d^3k_2}{(2\pi)^3}(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) \\
&\qquad \times aHfD_+^2(\eta)\left[1+\frac{\boldsymbol{k}_1\cdot\boldsymbol{k}_2}{k_1^2}\right]\delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2), \\
\theta^{\prime\,(2)}(\boldsymbol{k},\eta)+aH\theta^{(2)}(\boldsymbol{k},\eta)+\frac{3}{2}\Omega_{\rm m}(\eta)(aH)^2\delta^{(2)}(\boldsymbol{k},\eta) & \\
= -\int \frac{d^3k_1}{(2\pi)^3}\int \frac{d^3k_2}{(2\pi)^3}(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) & \\
\qquad \times (aHf)^2D_+^2(\eta)\left[\frac{\boldsymbol{k}_1\cdot\boldsymbol{k}_2}{k_1^2}+\frac{(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)^2}{k_1^2k_2^2}\right]\delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2). & \qquad (12.31)
\end{aligned}$$

In Exercise 12.3, you will fill in the intermediate steps we have skipped here, a very useful exercise for gaining proficiency in the real space–Fourier space correspondence. Notice that the complicated convolution integrals on the right-hand side do not depend on time; we can pull out the time-dependent factors involving $aHf$ and $D_+$.

In fact, we can make the equations even easier to solve by using the logarithm of the growth factor $\ln D_+$ as new time variable. Then

$$\delta_{\rm m}{}' = \frac{d\ln a}{d\eta}\frac{d\ln D_+}{d\ln a}\frac{\partial}{\partial\ln D_+}\delta_{\rm m} = aHf\frac{\partial}{\partial\ln D_+}\delta_{\rm m}.$$

While $\partial\delta^{(1)}/\partial\ln D_+ = \delta^{(1)}$ is simple enough, our goal is to derive the time evolution of $\delta^{(2)}$. Eq. (8.75) for the growth factor can be used to show that (Exercise 12.4)

$$\frac{d(aHf)}{d\eta} = (aH)^2\left(\frac{3}{2}\Omega_{\rm m}(\eta)-f(\eta)-f^2(\eta)\right). \qquad (12.32)$$

Finally, we define the scaled velocity divergence $\hat\theta \equiv \theta_{\rm m}/(aHf)$. With this, we obtain a simpler set of equations (again, Exercise 12.4):

$$\begin{aligned}
&\frac{d}{d\ln D_+}\delta^{(2)}(\boldsymbol{k},D_+)+\hat\theta^{(2)}(\boldsymbol{k},D_+) = D_+^2S_\delta(\boldsymbol{k}) \\
&\frac{d}{d\ln D_+}\hat\theta^{(2)}(\boldsymbol{k},D_+)+\left(\frac{3}{2}\frac{\Omega_{\rm m}(D_+)}{f^2(D_+)}-1\right)\hat\theta^{(2)}(\boldsymbol{k},D_+)+\frac{3}{2}\frac{\Omega_{\rm m}(D_+)}{f^2(D_+)}\delta^{(2)}(\boldsymbol{k},D_+) = D_+^2S_\theta(\boldsymbol{k}).
\end{aligned} \qquad (12.33)$$

Here, the time-independent source terms are given by

$$
\begin{aligned}
S_\delta(\boldsymbol{k}) &= \int \frac{d^3k_1}{(2\pi)^3} \int \frac{d^3k_2}{(2\pi)^3} (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) \\
&\qquad \times \left[1 + \frac{\boldsymbol{k}_1\cdot\boldsymbol{k}_2}{k_1^2}\right] \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2), \\
S_\theta(\boldsymbol{k}) &= -\int \frac{d^3k_1}{(2\pi)^3} \int \frac{d^3k_2}{(2\pi)^3} (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) \\
&\qquad \times \left[\frac{\boldsymbol{k}_1\cdot\boldsymbol{k}_2}{k_1^2} + \frac{(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)^2}{k_1^2 k_2^2}\right] \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2).
\end{aligned}
\tag{12.34}
$$

In the ΛCDM cosmology and dark energy cosmologies with similar expansion histories, it turns out that the quantity $\Omega_{\rm m}(\eta)/f^2(\eta)$ is very close to 1. Recall from Eq. (8.78) that the growth rate can be well approximated by $f(\eta) \simeq \left[\Omega_{\rm m}(\eta)\right]^{0.55}$. Thus, it is a good approximation (in practice, better than 1% in $\delta^{(2)}$, $\theta^{(2)}$), to set this ratio to unity. Then, the only terms in Eq. (12.33) that depend explicitly on time (via $D_+$) are the source terms. Let us then make the following power-law ansatz:

$$
\delta^{(2)}(\boldsymbol{k}, D_+) = A_\delta(\boldsymbol{k}) D_+^n; \quad \hat{\theta}^{(2)}(\boldsymbol{k}, D_+) = A_\theta(\boldsymbol{k}) D_+^n. \tag{12.35}
$$

Inserting this into Eq. (12.33) yields

$$
\begin{aligned}
n A_\delta D_+^n + A_\theta D_+^n &= D_+^2 S_\delta, \\
n A_\theta D_+^n + \frac{1}{2} A_\theta D_+^n + \frac{3}{2} A_\delta D_+^n &= D_+^2 S_\theta.
\end{aligned}
\tag{12.36}
$$

Clearly, for this to hold at all times $D_+$, we need $n = 2$. With this, solving for $A_\delta$ and $A_\theta$ yields

$$
\begin{aligned}
A_\delta(\boldsymbol{k}) &= \frac{5}{7} S_\delta(\boldsymbol{k}) - \frac{2}{7} S_\theta(\boldsymbol{k}), \\
A_\theta(\boldsymbol{k}) &= -\frac{3}{7} S_\delta(\boldsymbol{k}) + \frac{4}{7} S_\theta(\boldsymbol{k}).
\end{aligned}
\tag{12.37}
$$

Note that this is only one, the fastest-growing solution, but this is the one we are interested in anyway. Going back to conformal time $\eta$, we can thus write

$$
\begin{aligned}
\delta^{(2)}(\boldsymbol{k}, \eta) &= D_+^2(\eta) \int \frac{d^3k_1}{(2\pi)^3} \int \frac{d^3k_2}{(2\pi)^3} (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) \\
&\qquad \times F_2(\boldsymbol{k}_1, \boldsymbol{k}_2) \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2), \\
\theta^{(2)}(\boldsymbol{k}, \eta) = aHf\hat{\theta}^{(2)} &= -aHf D_+^2(\eta) \int \frac{d^3k_1}{(2\pi)^3} \int \frac{d^3k_2}{(2\pi)^3} (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k}-\boldsymbol{k}_1-\boldsymbol{k}_2) \\
&\qquad \times G_2(\boldsymbol{k}_1, \boldsymbol{k}_2) \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2),
\end{aligned}
\tag{12.38}
$$

where

$$F_2(\boldsymbol{k}_1,\boldsymbol{k}_2)=\frac{5}{7}+\frac{2}{7}\frac{(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)^2}{k_1^2k_2^2}+\frac{1}{2}\boldsymbol{k}_1\cdot\boldsymbol{k}_2\left(\frac{k_1}{k_2}+\frac{k_2}{k_1}\right),$$
$$G_2(\boldsymbol{k}_1,\boldsymbol{k}_2)=\frac{3}{7}+\frac{4}{7}\frac{(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)^2}{k_1^2k_2^2}+\frac{1}{2}\boldsymbol{k}_1\cdot\boldsymbol{k}_2\left(\frac{k_1}{k_2}+\frac{k_2}{k_1}\right). \tag{12.39}$$

We have symmetrized these kernels in $\boldsymbol{k}_1$, $\boldsymbol{k}_2$ for convenience, since they are integrated against a symmetric integrand in Eq. (12.38).

We have thus obtained a closed-form solution for the second-order density and velocity fields, given the linear density field at a reference epoch $\delta_0(\boldsymbol{k})$. This procedure can be straightforwardly continued to higher order. For example, the equation for $\delta^{(3)}$, $\hat{\theta}^{(3)}$ looks exactly like Eq. (12.33) on the left-hand side; the source terms on the right-hand side now involve products of $\delta^{(1)}$ and $\delta^{(2)}$, $\hat{\theta}^{(2)}$, and scale as $D_+^3(\eta)$. Approximating $\Omega_\mathrm{m}/f^2=1$ again, the equations can be integrated analytically leading to $\delta^{(3)}$, $\hat{\theta}^{(3)}\propto D_+^3$. This continues to any higher order, and the $n$th order solution can be written as

$$\delta^{(n)}(\boldsymbol{k},\eta)=D_+^n(\eta)\left[\prod_{i=1}^n\int\frac{d^3k_i}{(2\pi)^3}\right](2\pi)^3\delta_\mathrm{D}^{(3)}\left(\boldsymbol{k}-\sum_{i=1}^n\boldsymbol{k}_i\right)$$
$$\times F_n(\boldsymbol{k}_1,\cdots,\boldsymbol{k}_n)\delta_0(\boldsymbol{k}_1)\cdots\delta_0(\boldsymbol{k}_n),$$
$$\theta^{(n)}(\boldsymbol{k},\eta)=aHf\hat{\theta}^{(n)}=-aHfD_+^n(\eta)\left[\prod_{i=1}^n\int\frac{d^3k_i}{(2\pi)^3}\right](2\pi)^3\delta_\mathrm{D}^{(3)}\left(\boldsymbol{k}-\sum_{i=1}^n\boldsymbol{k}_i\right)$$
$$\times G_n(\boldsymbol{k}_1,\cdots,\boldsymbol{k}_n)\delta_0(\boldsymbol{k}_1)\cdots\delta_0(\boldsymbol{k}_n). \tag{12.40}$$

This trivially holds for $n=1$ (linear order) as well if we define $F_1=G_1=1$. Notice that the $n$th order density and velocity fields involve precisely $n$ powers of the linear matter density $\delta_0$, as anticipated in the discussion below Eq. (12.29). The kernels $F_n$, $G_n$ are fully symmetric polynomials in their arguments, and can be computed iteratively order by order (for convenient recurrence relations, see Bernardeau et al., 2002).

This very neat result allows us to explicitly calculate how structure in the universe evolves nonlinearly. There is in fact an intuitive representation of the perturbative expansion in terms of diagrams, as shown in Fig. 12.2, which is closely analogous to the Feynman diagrams of quantum field theory. The second-order density field $\delta^{(2)}$ is constructed by joining two instances of the initial (linear) density field with an $F_2$ kernel. Similarly, the $n$th order field is made by joining $n$ initial density fields with the $n$th order kernel $F_n$. The analogous rules hold for the expansion of the velocity divergence.

Most importantly, the perturbation-theory prediction Eq. (12.40) allows us to compute the statistics of the nonlinear, evolved density in terms of the statistics of the linear field $\delta_0(\boldsymbol{k})$. The power spectrum of $\delta_\mathrm{m}(\boldsymbol{k})$ can be written as

$$\left\langle\delta_\mathrm{m}(\boldsymbol{k},\eta)\delta_\mathrm{m}(\boldsymbol{k}',\eta)\right\rangle=\sum_{n,l=1,2,\cdots}^{n+l\ \text{even}}\left\langle\delta^{(n)}(\boldsymbol{k},\eta)\delta^{(l)}(\boldsymbol{k}',\eta)\right\rangle. \tag{12.41}$$

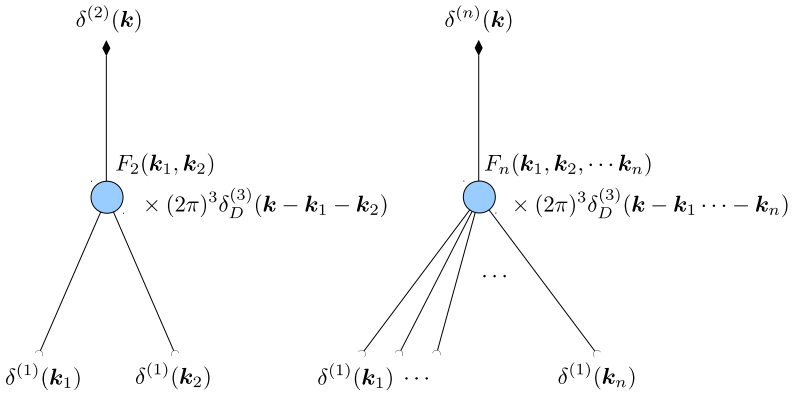



**FIGURE 12.2** Diagrammatic representation of the second-order density field $\delta^{(2)}$ (left) and the $n$th order density field (right). In each case, the final density field is connected to $n$ initial density fields by the interaction kernel $F_n$ (with $n = 2$ in the case of $\delta^{(2)}$). Analogous diagrams describe the velocity divergence $\theta^{(n)}$ in terms of kernels $G_n$. Here we suppress the time arguments for clarity.

Now, this result is not very practical, since we have to sum over infinitely many terms. In fact, perturbation theory makes sense only if we can truncate the sum after a finite number of terms, and the discarded terms are smaller than the ones we include. So let us look at the first three terms in the sum:

$$\begin{aligned}\left\langle \delta_{\rm m}(\boldsymbol{k},\eta)\delta_{\rm m}(\boldsymbol{k}',\eta)\right\rangle &= D_+^2(\eta)\left\langle \delta_0(\boldsymbol{k})\delta_0(\boldsymbol{k}')\right\rangle \\ &\quad + \left\langle \delta^{(2)}(\boldsymbol{k},\eta)\delta^{(2)}(\boldsymbol{k}',\eta)\right\rangle + 2\left\langle \delta^{(1)}(\boldsymbol{k},\eta)\delta^{(3)}(\boldsymbol{k}',\eta)\right\rangle + \cdots . \end{aligned} \tag{12.42}$$

The first line contains the linear power spectrum at time $\eta$. The terms in the second line make up the leading nonlinear correction to the matter power spectrum, i.e. the *next-to-leading order (NLO)* matter power spectrum. They can be expanded using the fact that $\delta_0$ is a Gaussian field (see Box 12.1); in fact we have already dropped terms that involve three fields $\delta_0$ in Eqs. (12.41)–(12.42), since they vanish.

## 12.1 Gaussian random fields

In cosmology, we usually compress the information in fields such as the matter density field into summary statistics, like the by-now familiar matter power spectrum. We have learned that the linear matter density $\delta_0$ is a Gaussian random field, a property inherited from the quantum fluctuations during inflation. Let us now define this more precisely. We begin in real space. A general Gaussian random field $\delta_0(\boldsymbol{x})$ with vanishing mean is completely specified by its two-point correlation function,

$$\langle \delta_0(\boldsymbol{x}_1)\delta_0(\boldsymbol{x}_2)\rangle = \xi(\boldsymbol{x}_1 - \boldsymbol{x}_2), \tag{12.43}$$

which could be isotropic, $\xi(\boldsymbol{r}) = \xi(|\boldsymbol{r}|)$, but it does not have to be (while $\xi(-\boldsymbol{r}) = \xi(\boldsymbol{r})$ has to hold by symmetry). The expectation value of three fields, and in fact any odd number of fields,

vanishes:

$$\langle \delta_0(\boldsymbol{x}_1)\delta_0(\boldsymbol{x}_2)\delta_0(\boldsymbol{x}_3)\rangle = 0. \tag{12.44}$$

The expectation value with four fields is nonzero, but completely determined by $\xi(\boldsymbol{r})$:

$$\begin{aligned}\langle \delta_0(\boldsymbol{x}_1)\delta_0(\boldsymbol{x}_2)\delta_0(\boldsymbol{x}_3)\delta_0(\boldsymbol{x}_4)\rangle = {}& \xi(\boldsymbol{x}_1-\boldsymbol{x}_2)\xi(\boldsymbol{x}_4-\boldsymbol{x}_3)+\xi(\boldsymbol{x}_1-\boldsymbol{x}_3)\xi(\boldsymbol{x}_4-\boldsymbol{x}_2)\\ &+\xi(\boldsymbol{x}_1-\boldsymbol{x}_4)\xi(\boldsymbol{x}_3-\boldsymbol{x}_2),\end{aligned} \tag{12.45}$$

where the three terms arise from the three distinct possibilities of combining the four fields into two pairs, which each yield a correlation function via Eq. (12.43). This expansion by pairing fields similarly works for any higher, even number of fields, and it is known as *Wick's theorem*. The Fourier-space counterparts to Eqs. (12.43)–(12.45) can be derived straightforwardly by taking the Fourier transform (we in fact highly recommend readers to go through these steps). We obtain

$$\left\langle \delta_0(\boldsymbol{k})\delta_0(\boldsymbol{k}')\right\rangle = (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}')P(\boldsymbol{k}), \tag{12.46}$$

where $P(\boldsymbol{k})$ is the Fourier transform of $\xi(\boldsymbol{r})$, and

$$\begin{aligned}\langle \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2)\delta_0(\boldsymbol{k}_3)\rangle &= 0\\ \langle \delta_0(\boldsymbol{k}_1)\delta_0(\boldsymbol{k}_2)\delta_0(\boldsymbol{k}_3)\delta_0(\boldsymbol{k}_4)\rangle &= (2\pi)^6\delta_{\rm D}^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_2)\delta_{\rm D}^{(3)}(\boldsymbol{k}_3+\boldsymbol{k}_4)P(\boldsymbol{k}_1)P(\boldsymbol{k}_3)\\ &\quad+(2\pi)^6\delta_{\rm D}^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_3)\delta_{\rm D}^{(3)}(\boldsymbol{k}_2+\boldsymbol{k}_4)P(\boldsymbol{k}_1)P(\boldsymbol{k}_2)\\ &\quad+(2\pi)^6\delta_{\rm D}^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_4)\delta_{\rm D}^{(3)}(\boldsymbol{k}_2+\boldsymbol{k}_3)P(\boldsymbol{k}_1)P(\boldsymbol{k}_2).\end{aligned} \tag{12.47}$$

The NLO contributions can be evaluated directly by inserting the solution Eq. (12.40), and using Wick's theorem Eq. (12.47). Again, the diagrammatic representation illustrates this formalism intuitively (Fig. 12.3): the power spectrum correlates two evolved density fields. Our goal is to connect them using their relation to the linear density fields shown in Fig. 12.2. So, we contract the instances of the linear density field in pairs, where each pair results in a linear power spectrum $P_{\rm L}$. The simplest way to connect is to just directly pair the final density fields. This is the leading, "tree-level" contribution, which is the linear power spectrum $P_{\rm L}(k)$. There are two ways to connect the evolved field using four linear fields, yielding two linear power spectra, which are the two contributions making up the next-to-leading order in Eq. (12.42). Analogous to the Feynman diagrams of field theory, there are precise rules underlying the diagrams (deriving these rules is left as an exercise to the field-theory-inclined reader), which offer an efficient shortcut to the underlying equations. Alternatively, one can go ahead and compute directly using Wick's theorem, which at this order is not much slower.

As you will derive in Exercise 12.5, the result is

$$\begin{aligned}P(k,\eta) &= P_{\rm L}(k,\eta)+P^{\rm NLO}(k,\eta)+\cdots,\\ P^{\rm NLO}(k,\eta) &= P^{(22)}(k,\eta)+2P^{(13)}(k,\eta),\end{aligned} \tag{12.48}$$

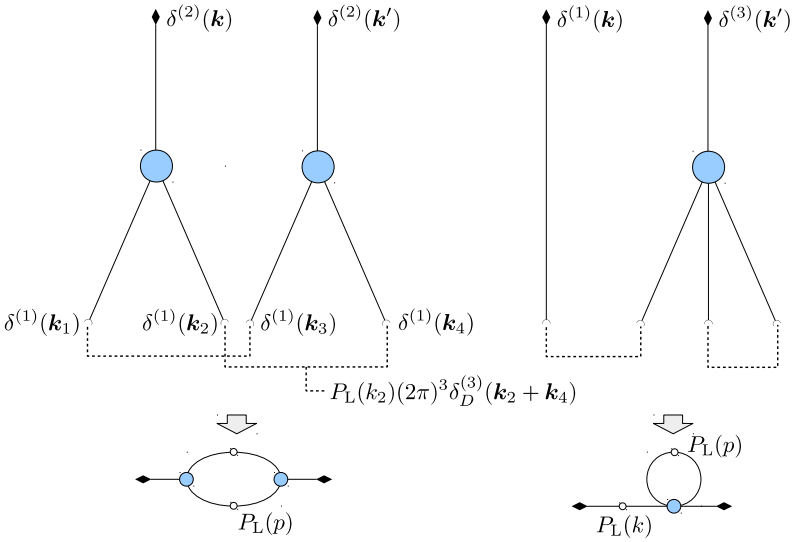



**FIGURE 12.3** Diagrammatic representation of the next-to-leading order contributions to the matter power spectrum: $\langle\delta^{(2)}(\boldsymbol{k})\delta^{(2)}(\boldsymbol{k}')\rangle$ (left) and $\langle\delta^{(1)}(\boldsymbol{k})\delta^{(3)}(\boldsymbol{k}')\rangle$ (right); we again suppress the time arguments for clarity. The upper diagrams show how these contributions can be calculated by connecting the linear density fields $\delta^{(1)}(\boldsymbol{k}_1), \cdots \delta^{(1)}(\boldsymbol{k}_4)$ appearing in the expansion of each nonlinear density field via the dashed lines (the kernels are the same as in Fig. 12.2 and are not labeled). By Wick's theorem, each connection yields a linear matter power spectrum and a Dirac delta. The lower diagrams introduce a more standard, and economical representation: now the connection of two linear fields is represented with an open circle, with each circle corresponding to a linear power spectrum. This representation makes it clear why these contributions are also called "1-loop" contributions. Each loop in a diagram corresponds to one integral over wavenumber (in the lower diagrams, $p$ denotes the loop wavenumber).

where

$$P^{(22)}(k,\eta) = 2\int \frac{d^3p}{(2\pi)^3}\left[F_2(\boldsymbol{p},\boldsymbol{k}-\boldsymbol{p})\right]^2 P_{\rm L}(p,\eta)P_{\rm L}(|\boldsymbol{k}-\boldsymbol{p}|,\eta),$$
$$P^{(13)}(k,\eta) = 3P_{\rm L}(k,\eta)\int \frac{d^3p}{(2\pi)^3}F_3(\boldsymbol{p},-\boldsymbol{p},\boldsymbol{k})P_{\rm L}(p,\eta). \tag{12.49}$$

Here, we have relabeled the wavenumbers $\boldsymbol{k}_i$ that are integrated over as $\boldsymbol{p}$. Notice that we have to go to third order to consistently derive the NLO correction to the matter power spectrum. The result is shown in Fig. 12.4. We see that on large scales (small $k$), $P^{\rm NLO}(k)$ is much smaller than the linear power spectrum. That is, nonlinear evolution is only a small correction to linear evolution. This is the regime where perturbation theory is useful, since we expect that higher-order terms in the expansion Eq. (12.48) are even smaller.

In fact, we can make this argument more precise. Notice that, as depicted in the bottom panel of Fig. 12.3, the NLO contributions in Eq. (12.49) both involve what in field theory is called a *loop*, an integral over wavenumber (or "momentum"). Since the linear matter power spectrum does not have a simple shape, this integral has to be performed numerically. In order to identify the relevant parameter controlling the relative size of the NLO

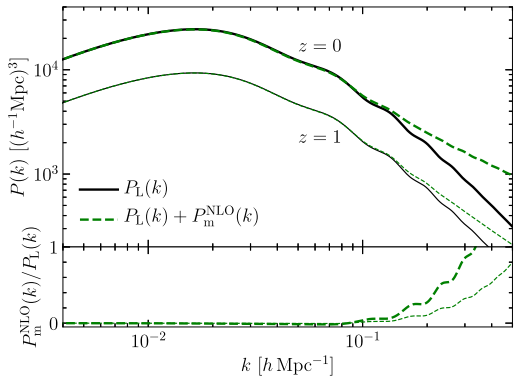



**FIGURE 12.4** Linear and next-to-leading order matter power spectrum [Eq. (12.48)] (top panel), at $z = 0$ (thick lines) and $z = 1$ (thin lines). The bottom panel shows the ratio of the NLO to linear power spectra. Perturbation theory is expected to break down when the NLO correction becomes of similar magnitude to the linear power spectrum itself, which in the fiducial cosmology happens for $k \approx 0.3\,h\,{\rm Mpc}^{-1}$ ($z = 0$) and $k \approx 0.6\,h\,{\rm Mpc}^{-1}$ ($z = 1$), respectively, close to $k_{\rm NL}$ in each case.

contribution compared to the linear matter power spectrum, we can use the fact that the perturbation-theory kernels are typically of order one. Then we can guess that this parameter is

$$\sim \int^k \frac{d^3p}{(2\pi)^3} P_{\rm L}(p) = \frac{1}{2\pi^2}\int_0^k p^2 dp\, P_{\rm L}(p), \tag{12.50}$$

which corresponds to the variance of the linear density field filtered on a spatial scale $R \sim 1/k$ [cf. Eq. (12.4)]. Perhaps you ask why we cut off the integral over $p$ at the scale $k$. The mathematical reason is that the perturbation-theory kernels in Eq. (12.48) are suppressed when $p \gg k$. The physical reason is that very small-scale perturbations to the matter density field do not influence the large-scale perturbations: the gravitational effect of a clump of matter, far away from the clump, only depends on its total mass, and is independent of how the mass is distributed within it. So, very roughly the fractional next-to-leading-order correction to the linear power spectrum is given by $\sigma^2_{R=k^{-1}}$. This becomes of order unity when $k \simeq k_{\rm NL}$, where recall we have defined the nonlinear wavenumber $k_{\rm NL}$ as the scale where the dimensionless linear matter power spectrum is equal to 1 (Sect. 8.1.1). Fig. 12.4 confirms this estimate. Notice that the regime where perturbation theory is valid extends to significantly smaller scales at redshift $z = 1$ compared to $z = 0$.

Another important effect of nonlinear evolution is that statistics involving an odd number of matter density fields no longer vanish. The leading example is the Fourier-space three-point correlation function, or *bispectrum*, which is given by

$$\langle \delta_{\rm m}(\boldsymbol{k}_1,\eta)\delta_{\rm m}(\boldsymbol{k}_2,\eta)\delta_{\rm m}(\boldsymbol{k}_3,\eta)\rangle = (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3) \times \left[2F_2(\boldsymbol{k}_1,\boldsymbol{k}_2)P_{\rm L}(k_1,\eta)P_{\rm L}(k_2,\eta) + 2\ \text{perm.}\right]. \tag{12.51}$$

This again follows from inserting Eq. (12.40), and using Wick's theorem Eq. (12.47) (Exercise 12.6). Note that the bispectrum is a function of three wavenumbers, and is nonzero only if these vectorially add up to zero, i.e. they form a closed triangle in Fourier space. The amplitude of the bispectrum in Eq. (12.51) displays a specific dependence on the shape of the triangle, which is characteristic of nonlinear gravitational evolution. Eq. (12.51) is only the leading-order result valid on large scales, and perturbation theory allows us to similarly calculate the next-to-leading order correction.

We now have all the tools we need to compute the statistics of the evolved matter density field in perturbation theory. However, before we move on, we should recall that, so far, we have actually done perturbation theory of the wrong equation: we have treated matter as an ideal fluid, whereas the real physical system is a collection of collisionless particles governed by the Vlasov equation. In particular, we have neglected the stress tensor $\sigma_{\rm m}^{ij}$ in Eq. (12.20). Fortunately, all is not lost: the solution is to treat matter as an *effective fluid* (Baumann et al., 2012). In practice, this works by expanding $\sigma_{\rm m}^{ij}$ in terms of the matter density field itself. Since we cannot predict $\sigma_{\rm m}^{ij}$ from within perturbation theory, we have to allow for free coefficients that must be determined by other means. The equation for $\boldsymbol{u}_{\rm m}$ involves only the gradient of $\sigma_{\rm m}^{ij}$, so the homogeneous part of the stress tensor is irrelevant. Hence, the leading relevant term is proportional to $\delta_{\rm m}$ and is given by

$$\sigma^{ij}_{\rm m,eff}(\boldsymbol{x},\eta) = \delta^{ij}\,\rho_{\rm m}(\eta)\,c^2_{s,\rm eff}(\eta)\delta_{\rm m}(\boldsymbol{x},\eta), \tag{12.52}$$

where $c^2_{s,\rm eff}$ is the effective sound speed squared. This notation makes sense: the diagonal part of the stress tensor corresponds to the pressure, and the sound speed $c_s^2 = \partial p/\partial\rho$ relates pressure perturbations to density perturbations. Note that this is not pressure in the usual sense as in a gas of collisional atoms. Instead, it corresponds to the effective gravitational action induced by small-scale perturbations. It is straightforward to integrate the Euler–Poisson system with this pressure term included. At linear order, this yields

$$\delta^{(1)}(\boldsymbol{k},\eta) = \left[1 - C_s^2(\eta)k^2\right] D_+(\eta)\delta_0(\boldsymbol{k}), \tag{12.53}$$

where $C_s^2(\eta)$ is a double time integral (weighted by the growth factor) over $c^2_{s,\rm eff}$. Notice how the effective pressure contribution is suppressed at small $k$, just as the NLO contribution we computed above. In fact, it is typically of similar order as the latter: on dimensional grounds, we expect that $C_s^2 \sim 1/k_{\rm NL}^2$, and simulation measurements confirm this. Thus, we can take into account the non-ideal nature of the effective fluid, i.e. the error we are making by approximating matter as a fluid, by performing an expansion of the stress tensor, with a single term being sufficient at the level of the power spectrum at NLO. The coefficient $C_s^2(\eta)$ cannot be predicted in perturbation theory. In order to determine it, we need to match to a solution of the actual underlying Vlasov–Poisson system. N-body simulations provide a means to achieve just that.

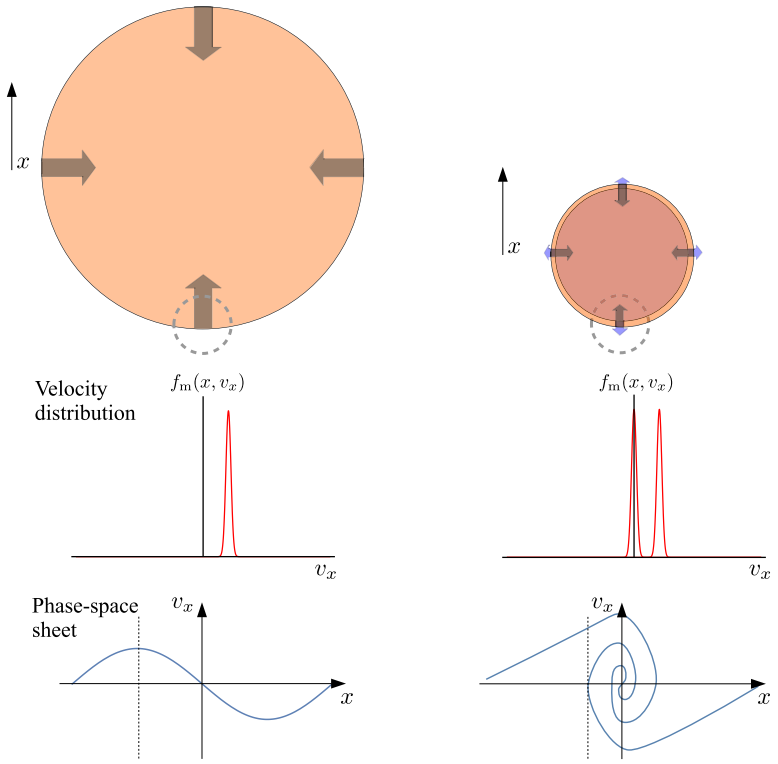



**FIGURE 12.5** Illustration of collapse at early times, where the velocity distribution is single-valued (left), and late times, in the multistreaming regime where $f_{\rm m}(\boldsymbol{x}, \boldsymbol{v}, t)$ has several peaks (right). The dynamics in the stage illustrated on the left can be described by an effective fluid, the ones on the right cannot. *Upper panels:* sketch of the configuration in real space. *Middle panels:* velocity distributions at the location of the dashed circles in the corresponding upper panel. *Lower panels:* phase-space distribution of matter. The distribution remains localized in a thin sheet. The vertical lines in each case indicate the location for which the velocity distribution is shown in the middle panels.

## 12.3 Simulations

In the previous section, we described how taking moments of the Vlasov equation leads to fluid equations for collisionless matter, which we were then able to solve perturbatively. However, the fluid equations do not correctly describe the evolution of nonlinear structure on small scales. We already mentioned this above, but let us study the issue in a bit more detail. Consider an overdense region that collapses under its own gravity (Fig. 12.5). Initially, the velocity of matter at the outer edge is single-valued (left middle panel). Eventually however, this shell encounters a shell that started at a smaller initial radius and already had time to pass through the origin; since dark matter is collisionless, a shell passes through the origin unimpeded. Thus, at the instant and location highlighted in the right panels, commonly called *shell crossing*, the velocity distribution now has two peaks (right middle panel), one corresponding to an infalling velocity from the outer shell, and another with

close to zero velocity from the inner shell. That is, for the location and time shown in the right panels, the outer shell is still on its first infall, while the inner shell has just reached $v_x = 0$ and is about to recollapse.

In the fluid treatment, on the other hand, two clouds of mass cannot pass through each other; instead, the pressure forces ultimately become important and the fluid would produce a shock. Mathematically, a fluid always has a single well-defined velocity $\boldsymbol{u}(\boldsymbol{x}, t)$ at any given point in space and time, so cannot describe the multivalued velocity (i.e. distribution function with several peaks) during shell crossing. This difference is explained by the fact that, in the fluid description, we have neglected the contribution from the stress tensor $\sigma^{ij}_{\rm m}$ as well as higher moments of the distribution function $f_{\rm m}$. On small scales, where shell crossing happens,[4] all moments of the distribution function become important.

What other means do we have to follow the evolution of collisionless matter? Let us go back to the Vlasov–Poisson system of Eq. (12.8):

$$\begin{aligned} &\frac{\partial f_{\rm m}}{\partial t} + \frac{\partial f_{\rm m}}{\partial x^j}\frac{p^j}{ma} - \frac{\partial f_{\rm m}}{\partial p^j}\left[Hp^j + \frac{m}{a}\frac{\partial \Psi}{\partial x^j}\right] = 0, \\ &\nabla^2\Psi = 4\pi G a^2\left[\int \frac{d^3p}{(2\pi)^3} f_{\rm m}(\boldsymbol{x}, \boldsymbol{p}, t) - \rho_{\rm m}(t)\right]. \end{aligned} \tag{12.54}$$

Our goal is to solve for $f_{\rm m}$, starting from cold initial conditions as given in Eq. (12.9):

$$f_{\rm m}(\boldsymbol{x}, \boldsymbol{p}, t) \xrightarrow{t\to 0} \frac{\rho_{\rm m}}{m}\left[1 + \delta_{\rm m}(\boldsymbol{x}, t)\right](2\pi)^3\delta^{(3)}_{\rm D}\left(\boldsymbol{p} - m\boldsymbol{u}_{\rm m}(\boldsymbol{x}, t)\right). \tag{12.55}$$

This initial condition states that matter occupies a thin sheet in phase space, with a unique single-valued velocity $\boldsymbol{u}_{\rm m}(\boldsymbol{x}, t)$ at each point in space. As $f_{\rm m}$ evolves under gravity, the velocity will no longer remain single-valued, as explained above, but matter will remain confined to a thin sheet in phase space, a consequence of the preservation of phase-space volume discussed in Sect. 3.2.1 (see the lower panels in Fig. 12.5). Whenever two clouds of matter pass through each other in physical space, this corresponds to a wrapping of the phase-space sheet.

N-body simulations proceed by discretizing this phase-space sheet and following its evolution numerically. A small element of the sheet has a well-defined position $\boldsymbol{x}$ and momentum $\boldsymbol{p}$. Since the motion of dark matter particles in this small region of phase space is described by the geodesic equation, so is that of the element of the phase-space sheet itself:

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{p^i}{ma}, \\ \frac{dp^i}{dt} &= -Hp^i - \frac{m}{a}\frac{\partial\Psi}{\partial x^i}. \end{aligned} \tag{12.56}$$

[4] The typical distance that a massive particle travels during the age of the universe is of order $10\,h^{-1}$ Mpc (see also Exercise 11.3), so we are safe from shell crossing on scales larger than this.

Mathematically, the non-relativistic geodesics are the *characteristics* of the collisionless Boltzmann equation. When integrating these equations numerically, it is convenient to use the "superconformal" momentum $\boldsymbol{p}_c \equiv a\,\boldsymbol{p}$. With this, the geodesic equation becomes

$$\begin{aligned}\frac{dx^i}{dt} &= \frac{p_c^i}{ma^2},\\ \frac{dp_c^i}{dt} &= -m\frac{\partial\Psi}{\partial x^i}.\end{aligned} \tag{12.57}$$

The advantage of $\boldsymbol{p}_c$ is that it is conserved in the absence of perturbations, i.e. when the gradient of $\Psi$ vanishes. Note that the coordinates $\boldsymbol{x}$ are comoving and thus include the Hubble expansion. Practitioners usually refer to the elements of the discretized phase-space sheet as "particles" for simplicity, and we will do so in the following as well. However, it is important to keep in mind that these *do not* stand for actual dark matter particles. Rather, they represent small elements of the dark matter distribution in phase space, which forms a thin sheet due to the cold nature of dark matter. For this reason, the mass $m$ of the particles (which we assume here is the same for all particles) is only a numerical parameter: it is determined by the total amount of matter in the simulation volume divided by the number of particles, so a higher-resolution simulation has more particles with correspondingly smaller $m$.

The basic sequence of an N-body simulation then proceeds as follows. Here, we describe the so-called leapfrog scheme where density and velocity are given at staggered times. So, we start with particle positions and velocities

$$\boldsymbol{x}^{(i)}(t) \qquad \text{and} \qquad \boldsymbol{p}_c^{(i)}(t-\Delta t/2), \tag{12.58}$$

where $\Delta t$ is the timestep and the superscript denotes the index of the particle. Typical simulations can have a billion particles or more, a number that is steadily growing with Moore's Law. We then

1. Compute the gravitational potential generated by the collection of particles, and take its gradient to obtain $\nabla\Psi(\boldsymbol{x},t)$ (see text below).
2. Change each particle's momentum ("kick") by

$$\boldsymbol{p}_c^{(i)}(t+\Delta t/2) = \boldsymbol{p}_c^{(i)}(t-\Delta t/2) - m\nabla\Psi(\boldsymbol{x}^{(i)},t)\Delta t. \tag{12.59}$$

3. Move each particle position ("drift") by

$$\boldsymbol{x}^{(i)}(t+\Delta t) = \boldsymbol{x}^{(i)}(t) + \frac{\boldsymbol{p}_c^{(i)}(t+\Delta t/2)}{ma^2(t+\Delta t/2)}\Delta t. \tag{12.60}$$

4. Repeat.

Notice that particle positions and momenta are offset by half a time step. This scheme ensures that the energy of each particle is conserved to high accuracy (the numerical error

in the energy is only of order $(\Delta t)^3$). The size of the time step $\Delta t$ used for each particle is often adjusted to the local acceleration $|\nabla\Psi|$ to guarantee accuracy in high-density regions. N-body simulations are typically performed in a cubic volume with periodic boundary conditions, so that particles exiting the volume on one side re-enter on the other side. This is appropriate if one aims to simulate a representative volume of a universe which is statistically homogeneous.

A crucial step in N-body simulations is the calculation of the potential gradient, i.e. the gravitational forces. Two main classes of algorithms exist: grid-based approaches, and the tree algorithm. In grid-based approaches, the mass of each particle is deposited on a 3D grid in order to obtain a smooth density field. The grid can either be of fixed resolution, or adaptively refined in regions with high particle density. Then, the Poisson equation (second line of Eq. (12.54)) is solved for $\Psi$ through a Fast Fourier Transform (FFT) or, in the case of adaptively refined grids, other fast numerical methods. Finally, the potential gradient is interpolated to each particle's position. In the tree algorithm one instead expands the force in multipoles, and keeps only the effect of the lowest multipoles from distant regions. Both classes of algorithms employ smoothing ("softening") of the force on small scales, in order to avoid direct particle-particle interactions; those would be unphysical since the "particles" do not in fact stand for physical particles.

The computational cost of both tree and adaptive grid codes scales roughly as $N \log N$ with the particle number $N$. This is to be contrasted with the cost of a direct summation of the force exerted on each particle by all other particles, which scales as $N^2$ and would be prohibitively expensive for simulations containing billions of particles.

Finally, we need to know how to set up the initial conditions of such a simulation. First, imagine distributing particles uniformly and at rest (in comoving coordinates), for example on a grid. A uniform density means that no potential gradient is produced, and the particles remain at rest. Now we slightly perturb the position of each particle. This leads to a density field with small perturbations. If we choose the velocities accordingly, we can generate a linear density field that evolves according to the growing mode of linear perturbation theory (see Exercise 12.7). In practice, the initial displacement field is generated in Fourier space, by drawing a random number for each Fourier mode from a Gaussian distribution with mean zero and variance given by the linear displacement power spectrum. Then, the displacement field is transformed back into real space.

To summarize: in order to perform an N-body simulation, we need (1) the matter density, determined by $\Omega_\mathrm{m}$; (2) the expansion history $a(t)$ or equivalently $H(a)$; and (3) the linear matter power spectrum from which to generate the initial conditions.

The final result of an N-body simulation is the collection of particle positions and velocities at different points in time ("snapshots"). By assigning particles to a grid, the density field can be computed, allowing for measurements of statistics such as the matter power spectrum (more advanced techniques actually use the fact that the particles represent a phase-space sheet). Moreover, one can search for gravitationally bound clusters of particles, *dark matter halos*, which indicate the typical locations of galaxies—although to actually see galaxies appear, one has to include gas and non-gravitational physics including

star formation. Finally, one can trace simulated light rays through the simulation volume, in order to make predictions for gravitational lensing.

Fig. 12.6 shows representations of the density field measured in a high-resolution N-body simulation. On large scales, we see that structure approaches homogeneity and isotropy. However, the density field is highly inhomogeneous on smaller scales, forming halos, filaments, and walls that surround large underdense regions (voids). Even inside bound halos, a hierarchy of substructure exists (lower panel). This hierarchical structure is a natural outcome of the amplitude of initial (linear) matter fluctuations as a function of scale in the concordance cosmology (Fig. 12.1), coupled with evolution under gravity. The small-scale fluctuations have the largest amplitude and hence collapse first to form bound structures. They subsequently become part of more massive halos, whereby their inner cores survive as substructures. Apart from hosting galaxies, halos can be interpreted as building blocks of the nonlinear structure. Hence, we will focus on them in the next section.

While the work of running and analyzing the output of N-body simulations is necessarily of numerical nature, the underlying problem and algorithms are simple and beautiful: solve the evolution of collisionless matter under gravity, starting from Gaussian initial conditions. Despite the simplicity of the problem, the outcome is remarkably rich and complex, as Fig. 12.6 shows.

Finally, we should mention the main caveat of the N-body simulations we have described here: they include only gravity, and consequently ignore baryonic pressure forces and the formation of stars and black holes. The latter two are particularly problematic: massive stars undergo supernova explosions, while black holes launch relativistic jets. These energetic phenomena can modify the matter distribution on small scales (both baryons and dark matter, which are coupled by gravity), an effect known as *feedback*. Since no cosmological simulation can resolve the formation of individual stars and black holes, these processes need to be treated with approximate "subgrid modeling." So, despite the power of modern simulations, there remains a theoretical uncertainty in how well we can predict the small-scale clustering of matter. This uncertainty currently is at the level of several percent at $k \simeq 1\,h\,\mathrm{Mpc}^{-1}$, and grows toward smaller scales. In Ch. 13, we will see that we can in fact measure the small-scale matter power spectrum using lensing, so that the limits in our understanding of baryonic feedback effects in the power spectrum need to be taken into account in cosmological constraints from gravitational lensing.

## 12.4 Dark matter halos

Bound structures are fairly straightforward to identify in N-body simulations: one searches for the very densest points in the matter distribution, and then evaluates which particles nearby are gravitationally bound. The latter condition can be evaluated in simulations, since we can compute the gravitational potential $\Psi_\mathrm{h}$ of the halo at a given particle's position and compare that with the particle's kinetic energy (a particle with velocity $v$ is bound if $v^2/2 < |\Psi_\mathrm{h}|$). On the other hand, particles are defined as "nearby" either if they are within

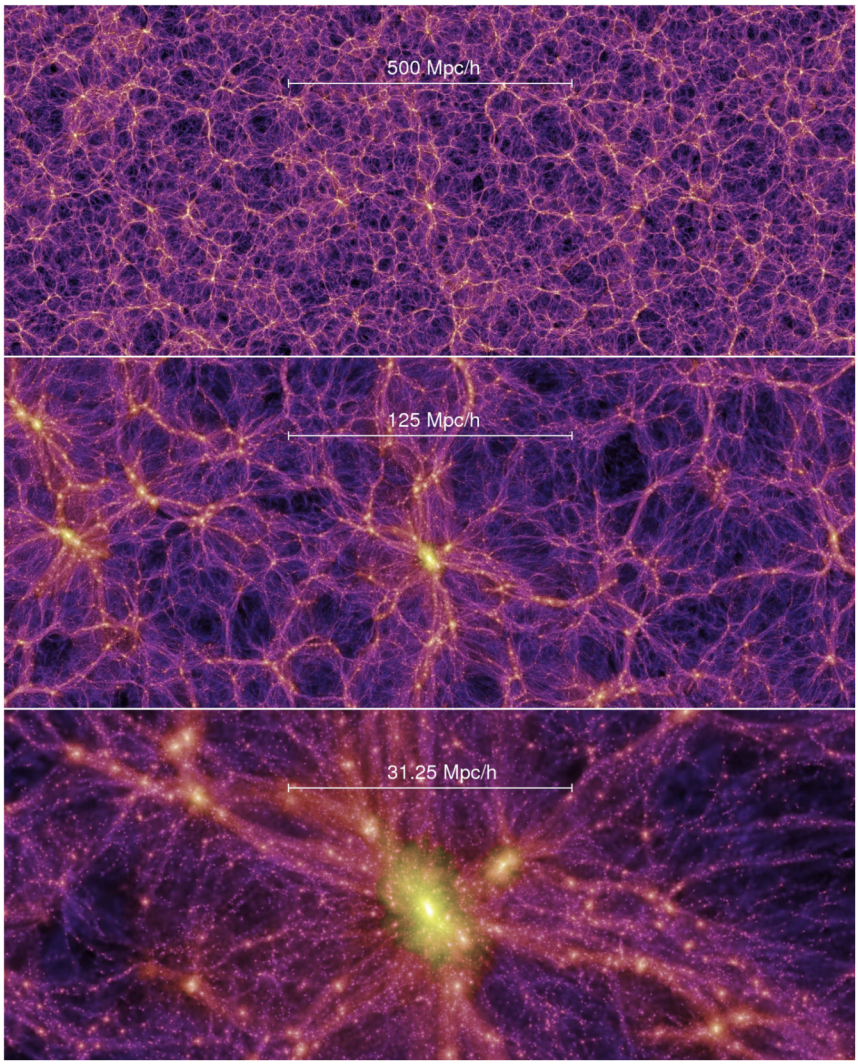



**FIGURE 12.6** Slices of width $15h^{-1}\,\mathrm{Mpc}$ through the density field at redshift zero in the *Millennium* N-body simulation which follows $10^{10}$ particles (i.e., phase-space elements). From top to bottom, the different panels zoom in to show the hierarchical nature of the matter distribution in a $\Lambda$CDM cosmology. The spatial scale is labeled in each panel. The color scale denotes density in logarithmic units. The simulations shown here are described in Springel et al. (2005).

a spherical region whose interior density is above some threshold ("spherical overdensity" algorithm), or if their nearest-neighbor distance to other halo particles is below a threshold value ("friends-of-friends" algorithm). Crucially, by definition any particle can be part of only a single halo. For both algorithms, the result is a catalog of halos with various masses, and various other properties, such as center-of-mass position and velocity.

But what is the significance of these structures and why are they called "halos"? The evidence for dark matter from the dynamics of stars and gas within galaxies, and of galaxies in clusters, goes back to the 1930s. Over the ensuing decades, it became clear that the dark component that is responsible for the additional gravitational potential has to be far more extended than the stars and the gas. Thus, the picture of a galaxy embedded in a much larger surrounding dark structure—the halo—was established. Even though N-body simulations, which only take into account gravity, do not form galaxies, the bound structures found in these simulations were soon identified with the halos hosting galaxies. Much evidence has since accumulated for this paradigm. Indeed, it rests on a fairly solid physical foundation. At high redshifts (around the epoch of reionization), the gas out of which stars eventually form cools most efficiently in dense regions, and this cooling allows it to collapse to sufficient density to trigger star formation. Hence, all galaxies are hosted by a dark matter halo of some mass, while the converse does not necessarily hold: there may well be low-mass halos which do not host a galaxy. Nevertheless, above a certain minimum halo mass, we believe that the majority of dark matter halos host at least one galaxy. Thus, if we know (or assume) how galaxies are distributed within halos as a function of halo mass, we can predict the abundance and clustering of galaxies based on a gravity-only N-body simulation—an enormous simplification over attempting to simulate the actual formation of galaxies.

Another application of halos is based on the fact that any particle can be part of only a single halo. If we further assume that all matter is enclosed in halos of some mass, we can build the entire matter density field out of the halo density field along with a model for their inner structure. This approach is referred to as the *halo model*, and we will return to it in Sect. 12.7.

### 12.4.1 Halo masses and profiles

Both of the applications mentioned above are aided by a fortunate fact about halos: while the detailed structure of individual halos is highly complex, their average properties are remarkably simple. To zeroth order, the properties of a halo at a given time are determined by a single number: its mass at that time. Before making use of this fact, we need to think about how to define a halo's mass. In simulations, one could strictly define it as the mass contained in all particles that are gravitationally bound. However, this definition is not particularly useful for connecting to observations, where we typically measure all visible or total matter within a given region centered on a halo. A more practical definition of the halo mass is to include all matter enclosed within a sphere around the halo center that encloses a fixed density, usually phrased in terms of a number $\Delta$ times the mean matter density. That is, one finds a radius $R_\Delta$ such that

$$\frac{M(<R_\Delta)}{4\pi R_\Delta^3/3} = \Delta \times \rho_{\rm m}(t_0), \tag{12.61}$$

where $R_\Delta$ is the comoving radius of the sphere (since simulations use comoving coordinates to follow the particles, it is convenient to use the same coordinates when analyzing

simulations as well). One then defines $M_\Delta \equiv M(< R_\Delta)$. If $\Delta$ is large, then almost all of the mass contained within this sphere is also bound to the halo. A typical choice is $\Delta = 200$, a number that emerges from an approximate semi-analytic calculation we describe in Sect. 12.4.2. In the literature you will thus frequently find numbers quoted for $R_{200}$ and the associated $M_{200}$. Note that $\Delta$ is sometimes also defined with respect to the critical density $\rho_{\rm cr}$, which corresponds in our convention to a $\Delta$ that is larger by a factor $1/\Omega_{\rm m}$ (we go back to our convention of $\Omega_{\rm m} = \Omega_{\rm m}(t_0)$ from now on).

Dark matter halos exhibit approximately universal spherically-averaged density profiles, as first demonstrated by Navarro et al. (1997). They proposed a simple fitting formula for average halo density profiles (*NFW profile*):

$$\rho_{\rm h}(r) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2}, \tag{12.62}$$

where $r_s$ is the *scale radius*. Notice that this profile is described by two parameters: $\rho_s$ and $r_s$. The scale radius is often parametrized by defining the *concentration* $c_\Delta \equiv R_\Delta/r_s$. Then, one can exchange the parameters $\rho_s$ and $r_s$ with the more practical parameters $M_\Delta$ and $c_\Delta$; you will derive this in Exercise 12.8. The concentration $c_\Delta$ is useful because it has been found to depend only weakly on halo mass. Using the density profile Eq. (12.62), it is then also possible to convert from one halo mass definition $(M_\Delta, R_\Delta)$ to a different one, $(M_{\Delta'}, R_{\Delta'})$. However, if you look at Eq. (12.62) carefully, you will realize that this profile cannot really describe bound halos at large radii: there, $\rho_{\rm h} \propto r^{-3}$, which means that the halo mass is logarithmically divergent. Instead, real halo profiles become steeper than the NFW form at large radii. This steepening typically happens for $r \gtrsim R_{200}$.

## 12.4.2 The halo mass function

The most important statistical property of halos is their abundance. This is typically phrased as the number density of halos $dn/d\ln M$ within an infinitesimal logarithmic mass bin $d\ln M$, known as the *halo mass function*, and can be directly measured in simulations. Moreover, we can indirectly estimate it observationally by using tracers of the underlying halos. For example, a large galaxy cluster can be identified with a massive halo, and an estimate for its mass obtained because the halo mass and the number of member galaxies are observed to be correlated with one another. We will discuss galaxy clusters, the observational counterparts to massive halos, in Sect. 12.5.

All rigorous results on dark matter halos need to be derived from full N-body simulations, and this also applies to the halo mass function. However, analytic approaches offer some neat insights, especially for the most massive and rarest halos which are the ones that host galaxy clusters. Because they are so rare, their formation can essentially be modeled as being in isolation (in contrast to lower-mass halos, which merge and are influenced by more massive neighbors).

The simplest halo formation scenario is the collapse of a uniform spherical region. The setup is illustrated in Fig. 12.7. Imagine a homogeneous universe at an early time $t_{\rm in}$ with matter density $\rho_{\rm m}(t_{\rm in})$, out of which we cut a sphere of mass $M$ (left panel). What radius in

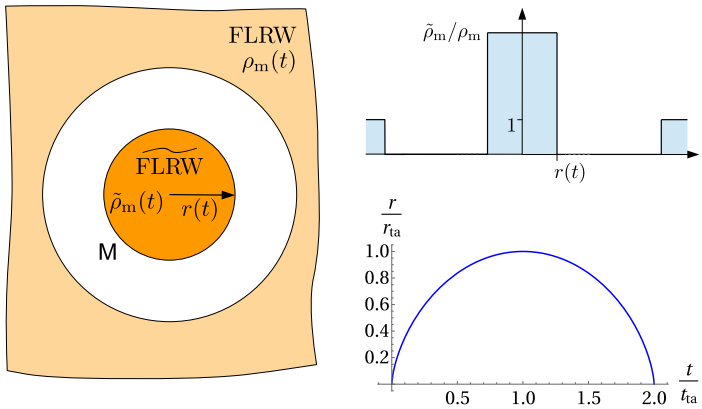



**FIGURE 12.7** Illustration of spherical collapse. Imagine cutting out a sphere of matter from the otherwise homogeneous universe, and compressing it slightly. This region will now begin to collapse, maintaining the enclosed mass $M$ and the spherical geometry (left, showing a plane projection; the upper right plot shows the density profile as a function of radius). The physical (rather than comoving) radius $r(t)$ then evolves as the scale factor of an FLRW universe with a larger density and positive curvature. The evolution of $r(t)$ is shown in the lower right plot. At $t = t_{\rm ta}$, $r(t)$ reaches turnaround and the region begins to collapse. Collapse $r = 0$ is reached at $t \simeq 2t_{\rm ta}$.

the unperturbed universe encloses a mass $M$? The mean density enclosed within a spherical region of *comoving* radius $R$ is

$$\rho_{\rm m}(t_0) = \Omega_{\rm m}\rho_{\rm cr} = \frac{M}{4\pi R^3/3}. \tag{12.63}$$

This relation defines the *Lagrangian radius* $R_L$ associated with the mass $M$:

$$R_L(M) = 1.40\,h^{-1}\,{\rm Mpc}\left(\frac{M}{10^{12}\,h^{-1}\,M_\odot}\right)^{1/3}, \tag{12.64}$$

where we used the value $\Omega_{\rm m} = 0.31$ of the fiducial cosmology. The Hubble constant drops out of this expression thanks to the units of $h^{-1}\,{\rm Mpc}$ and $h^{-1}\,M_\odot$. $R_L$ is equivalent to $R_\Delta$ with $\Delta = 1$. So if we imagine assembling a halo of mass $M \sim 10^{12}\,h^{-1}\,M_\odot$, which is roughly the Milky Way's halo mass, from the uniform matter density, we have to collect matter from within a comoving radius of about $1\,h^{-1}\,{\rm Mpc}$. Correspondingly, to form a massive halo hosting a galaxy cluster with $M \sim 10^{15}\,h^{-1}\,M_\odot$ we have to take matter from a region with radius $10\,h^{-1}\,{\rm Mpc}$. The name "Lagrangian radius" reflects what we just said: if we follow the constituent particles of the halo back to the initial conditions, then $R_L(M)$ is roughly the comoving size of the region that contains these particles, simply because the entire matter density field was close to the mean density initially. The comparison with Eq. (12.61) shows that $R_L(M_\Delta) = \Delta^{1/3}R_\Delta$ for halos identified with an interior density of $\Delta \times \rho_{\rm m}(t_0)$.

Now let us go back to the evolution of an initially slightly overdense homogeneous spherical region, with comoving radius $R_L$. The first crucial observation is that the mass $M$ is conserved, since none of the matter in the interior can escape to the outside, nor can

other material fall in; both of these are consequences of the spherically symmetric setup. From the point of view of an observer inside the overdense region making local measurements, this homogeneous region is indistinguishable from an FLRW universe with a higher background density $\tilde{\rho}_\mathrm{m}$ (marked by "$\widetilde{\mathrm{FLRW}}$" in Fig. 12.7). Let us then apply the second Friedmann equation (3.90) to this region:

$$\frac{\ddot{\tilde{a}}}{\tilde{a}} = -\frac{4\pi G}{3}\left[\tilde{\rho}_\mathrm{m} + 3\mathcal{P}\right], \tag{12.65}$$

where the dots refer to ordinary time derivatives as before and now $\tilde{\rho}_\mathrm{m}$ and $\mathcal{P}$ are the homogeneous density and pressure within the region. The only source of pressure $\mathcal{P}$ is the cosmological constant $\Lambda$, and this is the same as in the background universe. The *physical* (not comoving) radius $r(t)$ is proportional to the local scale factor within the overdensity, so we obtain

$$\frac{\ddot{r}}{r} = -\frac{4\pi G}{3}\left[\frac{M}{4\pi r^3(t)/3} - 2\rho_\Lambda\right]. \tag{12.66}$$

which becomes

$$\ddot{r}(t) = -\frac{GM}{r^2(t)} + \frac{8\pi G}{3}\rho_\Lambda r(t). \tag{12.67}$$

This is just the Newtonian equation of motion for a spherical mass of radius $r(t)$, augmented with the repulsive force due to the accelerated expansion caused by the cosmological constant or dark energy (note the opposite sign of this force, and that it increases with radius). Our initial condition for $r(t)$ at the early time $t_\mathrm{in}$, where the overdensity of the region is negligibly small, then is $r(t_\mathrm{in}) = a(t_\mathrm{in})R_L$, the factor of $a$ resulting from the conversion from comoving to physical radius. Similarly, $\dot{r}(t_\mathrm{in}) = \dot{a}(t_\mathrm{in})R_L = H(t_\mathrm{in})r(t_\mathrm{in})$, i.e. the region participates in the background Hubble flow.

Nothing stops us now from solving Eq. (12.67) numerically. However, if we drop the $\rho_\Lambda$ term, then the equation is solvable analytically (see Exercise 12.9). The solution is *parametric*, that is, radius and time are given as functions of a parameter $\theta$:

$$\begin{aligned} r(t) &= \frac{r_\mathrm{ta}}{2}(1-\cos\theta), \\ t &= \frac{t_\mathrm{ta}}{\pi}(\theta - \sin\theta). \end{aligned} \tag{12.68}$$

This solution is shown in the lower right panel of Fig. 12.7 and is straightforward to interpret. Initially, $r(t)$ is small and increasing since the region participates in the Hubble expansion; recall that $r$ is the physical radius. When $t$ approaches the *turnaround time* $t_\mathrm{ta}$, $\dot{r}$ goes through zero and becomes negative: the region begins to collapse. Collapse ($r = 0$) occurs precisely at $t = 2t_\mathrm{ta}$.

The parameters $r_\mathrm{ta}$ and $t_\mathrm{ta}$ depend on the size and initial overdensity of the region. Since we are able to calculate the statistics of the density field at early times, when linear perturbation theory applies, we would like to determine what initial overdensity is needed so that

collapse occurs at a given time $t$. One finds (see Exercise 12.9) that if the initial overdensity evolved forward using the linear growth factor, $\delta^{(1)}_{R_L}(t)$, exceeds a value of

$$\delta^{(1)}_{R_L}(t) > \delta_{\rm cr} = \frac{3}{5}\left(\frac{3\pi}{2}\right)^{2/3} \simeq 1.686, \tag{12.69}$$

then the region has collapsed by time $t$, i.e. it has reached $r = 0$. This is known as the *spherical collapse threshold*, and serves as a guide as to which regions in a given initial (linear) density field might collapse to form halos. Remarkably, the collapse threshold does not depend on the size and hence mass of the collapsing region. This is a consequence of the scale-free nature of a Euclidean matter-dominated universe. One can also estimate (Exercise 12.9) what the typical overdensity within $r(t)$ is when the spherical halo virializes, using the fact that virialization requires that the kinetic energy is $-1/2$ of the potential energy. Again, the result is independent of the size and mass of the region, and one finds $\Delta_{\rm vir} = 18\pi^2 \simeq 180$. This provides the motivation for choosing $\Delta = 200$ as the threshold density for defining halo mass and radius (given the rough nature of this approximation, it has become standard to choose the nearest round number for $\Delta$).

Finally, when including $\Lambda$, no closed-form solution for the spherical collapse exists. The equation is straightforward to integrate numerically though, and one finds that the effect of $\Lambda$ on $\delta_{\rm cr}$ and $\Delta_{\rm vir}$ is minor. The physical reason is that $\Lambda$ is subdominant during the early stages of collapse that happen in matter domination, while at late times, when $\Lambda$ becomes important, the collapsing region is already much denser than the background and has largely decoupled from the Hubble flow.

So, we have argued that the spherical collapse approximation is reasonable for very massive halos. How, then, to predict the abundance of galaxy clusters? This comes down to predicting the abundance of regions that collapse to form a halo following the condition Eq. (12.69). The basic insight comes from papers by Press and Schechter (1974) and Bond et al. (1991), and the resulting framework is called *extended Press–Schechter* or *excursion-set* theory. To understand the argument of these papers, consider the one-dimensional density field (blue wiggly line) in Fig. 12.8. There are regions with relatively large excursions in both the positive and negative directions. We are interested in the large excursions in the positive direction: it is these rare regions of large overdensity that collapse to form massive halos.

What is the fraction of space (in the initial conditions) that is contained in collapsed halos *above* mass $M$ at redshift $z$? Press and Schechter argued that, since the linear density field smoothed on a comoving scale $R_L(M)$ follows a Gaussian with zero mean and variance $\sigma(R_L, z)$, the volume fraction should simply be the integral over this Gaussian from the collapse threshold to infinity:

$$\begin{aligned} F_{\rm coll,PS}(M,z) &= 2 \times \frac{1}{\sqrt{2\pi}\sigma(R_L[M],z)} \int_{\delta_{\rm cr}}^{\infty} d\delta e^{-\delta^2/2\sigma^2(R_L[M],z)} \\ &= 2 \times \frac{1}{\sqrt{2\pi}} \int_{\delta_{\rm cr}/\sigma(R_L[M],z)}^{\infty} d\nu e^{-\nu^2/2}. \end{aligned} \tag{12.70}$$

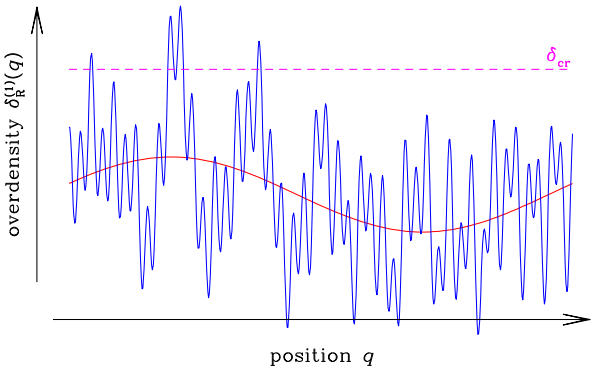



**FIGURE 12.8** Inhomogeneities as a function of 1D position. Shown is the initial, linear density field including a long-wavelength perturbation (red (mid gray in print version)) and the spherical collapse threshold $\delta_{\rm cr}$ [Eq. (12.69)]. From Desjacques et al. (2018).

Notice that the integral only depends on the ratio $\delta_{\rm cr}/\sigma(R_L[M], z)$. The factor of 2 was introduced by Press and Schechter (1974) as an ad hoc factor in order to recover the correct normalization. In particular, the expectation is that, as $R \to 0$, the variance diverges and hence all of matter should be contained in a collapsed structure (even though spherical collapse will not describe the formation of these low-mass objects correctly); that is, we expect that

$$\lim_{M\to 0} F_{\rm coll,PS}(M, z) = 1. \tag{12.71}$$

Without the prefactor 2 in Eq. (12.70), we would instead obtain $1/2$. The "fudge factor" was subsequently explained rigorously by Bond et al. (1991), who introduced the excursion-set formalism.

We now need to transform the collapsed fraction into the halo mass function. The halo mass function is simply the mean number density of matter multiplied by the fraction of matter that has collapsed into a halo of mass $M$. Therefore,

$$\frac{dn(M, z)}{d\ln M} = \frac{\rho_{\rm m}(t_0)}{M}\left|\frac{dF_{\rm coll,PS}}{d\ln M}\right|. \tag{12.72}$$

The factor $1/M$ in front comes about from converting the *mass density* in halos of mass $M$ (which is the equivalent of a volume fraction in Lagrangian space) to a *number density*. Using Eq. (12.70), the result is

$$\frac{dn(M, z)}{d\ln M} = \frac{\rho_{\rm m}(t_0)}{M} f_{\rm PS}\left(\frac{\delta_{\rm cr}}{\sigma(M, z)}\right)\left|\frac{d\ln\sigma(M, z)}{d\ln M}\right|, \quad f_{\rm PS}(\nu) = \sqrt{\frac{2}{\pi}}\nu e^{-\nu^2/2}, \tag{12.73}$$

where we have abbreviated $\sigma(M, z) \equiv \sigma(R_L[M], z)$, as is standard in the literature.[5] Massive halos have $\sigma(M, z) \ll \delta_{\rm cr}$, since the variance is small for large smoothing scales (Fig. 12.1).

[5] Note that, in the literature, our $f_{\rm PS}(\nu)$ is also frequently defined with a factor $\nu$ outside, i.e. $f_{\rm PS}(\nu) \to \nu f_{\rm PS}(\nu)$.

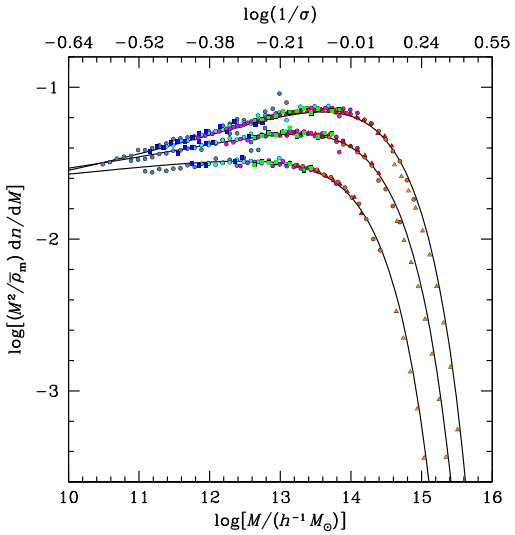



**FIGURE 12.9** Halo mass function (number density of halos per logarithmic mass interval), multiplied by mass, as measured in a large suite of N-body simulations (points). The line shows a parametrization of the form of Eq. (12.73) with $f_{\rm PS}(\nu)$ replaced with a fitting function which includes an exponential suppression at high $\nu$. The three sets of points show results for different halo mass definitions $M_\Delta$, with $\Delta = 200, 800$, and $3200$ (from top to bottom). From Tinker et al. (2008).

This corresponds to $\nu \gg 1$, and we see that the abundance of such halos is exponentially suppressed. This result is confirmed when counting halos in N-body simulations, as shown in Fig. 12.9. Within this toy model, it is easy to understand this fact: massive halos arise from rare upward fluctuations of the initial density field, which are suppressed by a $e^{-\nu^2/2}$ factor since the initial density field is Gaussian.

While Eq. (12.73) is only a very rough approximation to the mass function found in simulations, fitting formulas have been proposed that replace $f_{\rm PS}(\nu)$ with a more general function $f(\nu)$. With such a function, Eq. (12.73) remains fairly accurate for a range of masses, redshifts, and even different cosmologies (see the lines in Fig. 12.9, which covers five orders of magnitude in halo mass). Notice that with a general fitting function, the collapse threshold $\delta_{\rm cr}$ is no longer directly connected with spherical collapse. Nevertheless, spherical collapse and the excursion-set argument provide a reasonable physical picture of how the observed halo mass function comes about.

## 12.5 Galaxy clusters

Counting galaxy cluster continues to be an area of great fascination and promise for cosmology. First, we think we understand the theory fairly well: large clusters of galaxies are the most massive virialized structures in the universe, and extremely rare objects. It is thus physically well motivated to associate them with equally rare high-mass halos. We can therefore predict the abundance of halos as a function of mass, both analytically, and

more accurately, numerically, as we have discussed above. Second, the potential payoff is significant: since the high-mass end of the halo mass function depends exponentially on the parameter $\delta_{\rm cr}/\sigma(M)$, we expect that measuring the abundance of massive clusters will lead to very tight constraints on $\sigma(M)$, the amplitude of the matter power spectrum.

So let us turn to the practical aspects of cluster cosmology. There are two main challenges in this field. One class of difficulties is associated with finding clusters. The second difficulty then is to measure the mass of individual clusters. Sophisticated algorithms have been developed to find dense accumulations of galaxies in galaxy surveys, which is referred to as "optical identification." However, they can also be identified as luminous thermal X-ray sources, and through their specific imprints on the CMB "backlight" via the Sunyaev–Zel'dovich effect (Sect. 11.3).

Let us start with the first challenge and go through these techniques in a bit more detail. As we have learned, galaxy clusters form through the hierarchical merging of smaller virialized structures. Each of these smaller structures might host one or several galaxies. The end result is observed as a collection of galaxies mostly held together by the gravitational potential generated by the dark matter. It is thus physically reasonable to expect that the number of galaxies in an optically identified cluster, usually referred to as *richness*, is positively correlated with its mass: the more galaxies, the more matter and hence the higher the mass.

By far not all of the baryonic matter in clusters is contained in the galaxies however. Most of the baryons which joined the cluster during the gravitational collapse along with the dark matter are in the form of diffuse gas. During the process of virialization, this gas is heated to enormous temperatures (strictly, it is a plasma, since all light atoms are stripped of their electrons); so high, in fact, that its thermal radiation peaks in the *X-ray band*, making galaxy clusters some of the hottest objects in the universe. This fact unexpectedly turned the ROSAT X-ray satellite into a cluster finder. The recently launched eROSITA instrument will continue this effort through an all-sky survey with much increased sensitivity. The main observables are the temperature, inferred from the X-ray spectrum, and total X-ray luminosity of each cluster.

The free energetic electrons in galaxy clusters also produce the Sunyaev–Zel'dovich (SZ) effect (Sect. 11.3): they up-scatter CMB photons to higher energies via the inverse-Compton process. This leads to the characteristic *$y$-type distortion* of the observed CMB frequency spectrum in the direction of clusters. This distortion can be measured by observing the CMB at different frequencies, allowing for a separation from the primary CMB fluctuations (which are perfectly black-body). As we have seen in Sect. 11.3, the $y$ parameter is proportional to $n_e T$, which, by the ideal gas law, is the pressure of the gas. Clusters, with their high temperatures and large amount of free electrons, produce large SZ signals.

One of the unique properties of the thermal SZ effect is that it is very weakly dependent on the distance to the cluster. In contrast, both optical identification and X-rays rely on the direct detection of light from the cluster, which becomes increasingly difficult at large distances. On the other hand, the latter techniques have an advantage at low redshifts. This is apparent from Fig. 12.10, which shows a scatter plot in the mass-redshift plane of clusters

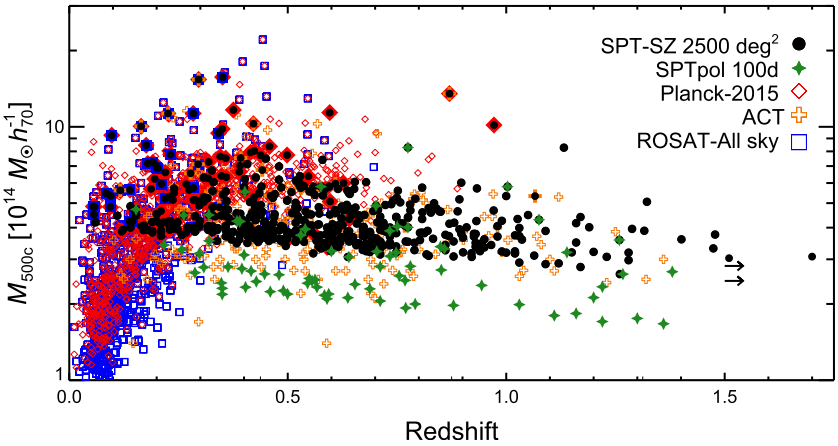



**FIGURE 12.10** Scatter plot of clusters detected using X-rays (ROSAT All sky X-ray survey) and the SZ effect (from 3 CMB experiments: South Pole Telescope (SPT-SZ, SPTpol), Planck, and Atacama Cosmology Telescope (ACT)) as a function of estimated mass ($M_\Delta$ with $\Delta = 500/\Omega_{\rm m}$, $y$-axis) and redshift ($x$-axis). The different distribution in the mass-redshift plane of the different cluster samples is evident: X-rays surveys are most sensitive to low-redshift clusters, while the SZ effect extends to much larger redshifts, but has less sensitivity to low masses. This plot is an updated version of that shown in Bleem et al. (2015).

detected using different techniques. At low redshifts, the X-ray identification typically has a lower mass threshold than SZ. The opposite holds at higher redshifts. Note that the overall number of clusters detected also depends on the size of the survey (for example, Planck and ROSAT observed the entire sky, while ACT and SPT cover only a small fraction).

So, X-ray and thermal SZ measurements allow us to get a handle on the temperature of the diffuse gas (or plasma) in clusters. How is this temperature related to its mass? Suppose a cluster has virialized so that its kinetic energy is equal to minus one half its potential energy. Suppose also that the cluster is spherical and of uniform density with radius $R_{\rm vir}$. The gravitational potential energy is then equal to $-3GM^2/5R_{\rm vir}$. Then,

$$\frac{1}{2}M\left\langle v^2\right\rangle = \frac{3}{10}\frac{GM^2}{R_{\rm vir}}, \tag{12.74}$$

where $\langle v^2\rangle$ is the velocity dispersion of the matter (dark matter and gas, assumed to be the same in this simple toy model). Assuming that the gas is dominated by hydrogen, ideal gas thermodynamics tells us that $\langle v^2\rangle/2 = (3/2)(kT/m_p)$, where $m_p$ is the proton mass. Further, we already know the overdensity that the cluster should roughly have, namely $\Delta = \Delta_{\rm vir} \approx 200$. This allows us eliminate the radius. The temperature can now be expressed in terms of the total mass of the system,

$$T = \frac{m_p}{5}\left[GMH_0\sqrt{\frac{\Delta_{\rm vir}\Omega_{\rm m}}{2}}\right]^{2/3}. \tag{12.75}$$

We invert this to get

$$M = 1.38 \times 10^{14}\, h^{-1}\, M_\odot \left(\frac{T}{\mathrm{keV}}\right)^{3/2} \left(\frac{\Delta_{\mathrm{vir}}}{200}\right)^{-1/2}. \tag{12.76}$$

Notice that, yet again, the Hubble constant drops out if we phrase the mass in terms of $h^{-1}\, M_\odot$. Thus, the temperature of the ionized gas in clusters tells us about their mass. It is important to keep some major caveats in mind, however. First, clusters do not have uniform densities; knowledge of the true density profile is important to relate the temperature to the mass. Second, we have assumed perfect thermalization of the gas, and have completely ignored any turbulent or bulk flows. Clusters are dynamically young objects and can thus be expected to have incompletely virialized. Third, we have neglected the effect of cooling due to the emitted thermal radiation, and of heating due to feedback from the galaxies residing in the clusters.

So, let us take stock. We have found several methods that allow us to identify massive collapsed objects in the universe, based on some observable which is related to the cluster mass. Now, each of these experimental techniques selects clusters on properties that are specific to each observable. That is, a cluster optically identified in a galaxy survey might not be detected as X-ray source, or vice versa. What we can predict however, is the abundance of massive virialized structures at fixed mass.

This leads us to the second set of difficulties, which revolves around determining the mass of a given cluster. We do not have a good theoretical handle on how the richness (number of galaxies) is related to the cluster mass. For X-rays and the Sunyaev–Zel'dovich distortion, we have some handle on the scaling with mass, but these are affected by numerous uncertainties as mentioned above. Fortunately, there is one technique which measures the mass fairly directly and model-independently: gravitational lensing (Ch. 13). The massive gravitational potential well of the cluster distorts light rays from more distant background galaxies. This effect can be detected statistically by measuring the shapes of a large number of background galaxies. In fact, the clusters also distort the CMB through lensing, an effect which has recently been measured for the first time. Weak lensing is now our best tool for calibrating cluster masses.

In most cases however, due to the limited signal-to-noise of the lensing signal, we do not have direct mass measurements for each individual cluster. The goal is then to calibrate the relation of the mass indicator, so for example richness, X-ray luminosity or SZ signal, to the mass itself. This statistical relation is referred to as *mass–observable relation*. Both the mean relation and the scatter about the mean relation are important quantities, since we need to quantify how likely a cluster of a given mass is to be included in the observed sample. This selection efficiency is a crucial ingredient in connecting the theoretical prediction to the observations. Of particular importance is a phenomenon known as *Malmquist bias*: since massive clusters are exponentially rare, it is much more likely that a cluster of lower mass is wrongly included in the sample due to an upward fluctuation in the mass indicator at fixed mass, than a higher-mass cluster is to be excluded due to a negative fluctuation.

The scatter in the mass–observable relation can thus significantly change the observed abundance of clusters, and needs to be modeled carefully.

## 12.6 Galaxy clustering and bias

In Ch. 11, we learned that a wealth of information can be extracted from the clustering of galaxies on large scales. We just had to make two simple but very important assumptions: 1) the galaxy overdensity $\delta_{\rm g}$ is linearly proportional to the matter overdensity, related by a bias factor $b_1$, and has a scale-independent noise contribution; 2) velocities of galaxies are the same on large scales as those of matter. We left open the question of why we can make these simple assumptions for objects as complicated as galaxies. They cannot possibly be entirely accurate, so what is the error we are making?

One approach to answer these questions would be to try to fully simulate the formation of galaxies, and measure the power spectrum of simulated galaxies. Unfortunately, simulating a realistic galaxy density field is *much* more difficult than obtaining an accurate matter density field. For matter, we argued that baryonic effects are relatively small and restricted to small scales. This does not hold for galaxies, since they are composed of baryons. As an example, small changes in the subgrid modeling can strongly affect both the abundance and clustering of simulated galaxies selected, for example, on simulated luminosity (and galaxies in actual surveys are often selected based on more complicated properties that are even harder to simulate than total luminosity). So, we need to somehow parametrize our ignorance of the galaxy formation process.

In the bulk of this section, we will pursue a perturbation-theory approach, which is what the results of Ch. 11 are based on, in which case the bias $b_1$ and the noise amplitude capture our ignorance on large scales. Making use of both perturbation theory (Sect. 12.2) and the toy model of halo formation we developed in Sect. 12.4.2, we will derive how this fact comes about and what the error is that we are making. At the end of the section, we will briefly discuss other approximate approaches based directly on simulations.

Let us begin with the bias of halos, based on what we learned about halo formation in the previous section. To compute the bias we need to predict $\delta_{{\rm h},\ell}$ in the presence of a large-scale overdensity in the matter, denoted as $\delta_\ell$. In Sect. 12.4.2, we argued that the number density of these halos is proportional to the fraction of regions in the initial density field that are above a critical threshold $\delta_{\rm cr}$. We can tweak this ansatz to account for a long-wavelength density perturbation $\delta_\ell$ (see the red (light gray in print version) line in Fig. 12.8). From the point of view of the collapsing regions, whose radius is much smaller than the wavelength of this perturbation, this just corresponds to adding or subtracting (depending on the sign of $\delta_\ell$) a uniform matter component on top of the small-scale fluctuations. This means that, in terms of the initial, linear density field, all the regions move closer to or further away from the collapse threshold by an amount $\delta_\ell$. That is, the collapse criterion Eq. (12.69) is modified to

$$\delta_R^{(1)}(\boldsymbol{x},t) > \delta_{\rm cr} - \delta_\ell^{(1)}(\boldsymbol{x},t), \tag{12.77}$$

where we assume linear evolution of the long-wavelength perturbation. Now we can compute what the expected number density of halos is by substituting Eq. (12.77) into Eq. (12.73):

$$\begin{aligned}\left.\frac{dn}{d\ln M}\right|_{\delta_\ell} &= \frac{\rho_{\rm m}(t_0)}{M} f_{\rm PS}\left(\frac{\delta_{\rm cr}-\delta_\ell^{(1)}}{\sigma(M,z)}\right)\left|\frac{d\ln\sigma(M,z)}{d\ln M}\right| \\ &\approx \left.\frac{dn}{d\ln M}\right|_0 \left[1-\frac{d\ln f_{\rm PS}}{d\nu}\frac{1}{\sigma(M,z)}\delta_\ell^{(1)}\right]_{\nu=\delta_{\rm cr}/\sigma(M,z)}, \end{aligned} \tag{12.78}$$

where in the second line we have expanded to first order in $\delta_\ell^{(1)}$, and the prefactor is the mean halo mass function. We then obtain for the fractional perturbation in the halo number density

$$\delta_{{\rm h},\ell}^{(1)}(\boldsymbol{x},t) = \frac{dn/d\ln M|_{\delta_\ell}}{dn/d\ln M|_0} - 1 \equiv b_1(M,z)\delta_\ell^{(1)}(\boldsymbol{x},t), \tag{12.79}$$

where the bias is defined as the coefficient relating the halo overdensity to the matter overdensity, so that

$$b_1(M,z) = -\left.\frac{1}{\sigma(M,z)}\frac{d\ln f_{\rm PS}(\nu)}{d\nu}\right|_{\nu=\delta_{\rm cr}/\sigma(M,z)}. \tag{12.80}$$

Plugging in the Press–Schechter form of the mass function, we obtain

$$b_1^{\rm PS}(M,z) = \left.\frac{\nu^2-1}{\delta_{\rm cr}}\right|_{\nu=\delta_{\rm cr}/\sigma(M,z)}. \tag{12.81}$$

This derivation of the bias is known as *peak-background split* argument. It was first made by Kaiser (1984), and can be justified rigorously using the same "separate universe" argument we made for the spherical collapse (Sect. 3 in Desjacques et al., 2018). Moreover, we do not have to assume the Press–Schechter expression $f_{\rm PS}(\nu)$ to compute the bias, but can insert more accurate parametrizations such as the one shown in Fig. 12.9.

We see that the bias coefficient becomes large if $\nu \gg 1$, i.e. for rare high-mass halos. This means that these halos cluster much more strongly than matter. The reason for this behavior becomes clear when looking at Fig. 12.8 carefully: there are many more peaks in the density field that are above the threshold $\delta_{\rm cr}$ when $\delta_\ell > 0$ (e.g., 3 in this case) than when $\delta_\ell < 0$ (zero), despite the fact that $\delta_\ell$ is not that large. The abundance of rare peaks thus reacts much more sensitively to a perturbation in the matter density than matter itself, which is in one-to-one correspondence with the fact that $b_1$ is much larger than 1.

We have thus found that the halo density perturbation is proportional to the matter density perturbation, with a proportionality constant $b_1$. Our derivation is valid since it focused on the effect of a very large-scale density perturbation, and we know from Sect. 12.2 that sufficiently large-scale perturbations can be treated as linear. How do we generalize this to smaller-scale perturbations, and what is the error we are making by restricting to a linear bias relation?

One approach is to follow the toy model that led us to Eq. (12.70) and Eq. (12.73): we assume that halos correspond to regions in the initial conditions that are above the collapse threshold $\delta_{\rm cr}$ when smoothed on the Lagrangian radius of the halo. Now, the correlation function of regions above the threshold $\delta_{\rm cr}$ at the separation $r$ is defined as the excess probability of finding a region above threshold at a distance $r$ from another region above threshold:

$$\xi_{\rm thr}(r) = \frac{p\left(\delta_R^{(1)}(\boldsymbol{x}+\boldsymbol{r}) > \delta_{\rm cr}, \delta_R^{(1)}(\boldsymbol{x}) > \delta_{\rm cr}\right)}{[p(\delta_R^{(1)}(\boldsymbol{x}) > \delta_{\rm cr})]^2} - 1. \tag{12.82}$$

Here and in the following, we again suppress the time arguments for clarity. Since the linear density field follows a multivariate Gaussian distribution, all of these probabilities can be written down analytically. In Exercise 12.10 you will perform this calculation, and show that the result can be written as a series expansion

$$\xi_{\rm thr}(r) = (b_1^{\rm thr})^2 \xi_R^{(1)}(r) + \frac{1}{2}(b_2^{\rm thr})^2 [\xi_R^{(1)}(r)]^2 + \cdots, \tag{12.83}$$

where $\xi_R^{(1)}$ is the correlation function of the linear matter density field smoothed on the scale $R$, and the dots stand for higher-order terms that involve three and more powers of the correlation function. $b_1^{\rm thr}$ is analogous to the linear bias coefficient we have derived above: the correlation function of regions above threshold is proportional to that of matter. The second term in the expansion involves a new bias parameter, the second-order bias $b_2^{\rm thr}$ (it corresponds to the coefficient we would obtain if we expanded to second order in $\delta_\ell$ in Eq. (12.78); see Exercise 12.11). The precise values of these coefficients are less important than the form of the terms we see in Eq. (12.83): if we work on sufficiently large scales $r$ so that $\xi_R^{(1)}(r) \ll 1$, then the higher-order bias terms are small corrections to linear bias. This justifies our linear bias treatment in Ch. 11.

The simple thresholding picture will not describe the actual galaxies whose power spectrum we measure observationally, but it gives us useful hints. Using the techniques we learned in Sect. 12.2, it is straightforward to obtain the Fourier-space version of Eq. (12.83):

$$\begin{aligned} P_{\rm g,thr}(k) = {} & (b_1^{\rm thr})^2 P_{\rm L}(k) W_R^2(k) \\ & + \frac{1}{2}(b_2^{\rm thr})^2 \int \frac{d^3p}{(2\pi)^3} P_{\rm L}(p) W_R^2(p) P_{\rm L}(|\boldsymbol{k}-\boldsymbol{p}|) W_R^2(|\boldsymbol{k}-\boldsymbol{p}|) \\ & + \cdots, \end{aligned} \tag{12.84}$$

where $W_R(k)$ is the filtering kernel in Fourier space (see Eq. (12.4)). We can set this kernel to 1 on large scales, i.e. when $k \ll 1/R$. Comparing Eq. (12.84) to Eq. (12.49), we see that the second-order bias contribution is of similar form, and hence of the same order (if $b_2^{\rm thr}$ is of order unity), as the next-to-leading order contribution $P^{(22)}(k)$ to the matter power spectrum in perturbation theory. This suggests that we can incorporate the bias expansion for galaxies into perturbation theory, by expanding the galaxy density perturbation in analogy

to Eq. (12.29):

$$\delta_{\rm g}(\boldsymbol{x},\eta)=\delta_{\rm g}^{(1)}(\boldsymbol{x},\eta)+\delta_{\rm g}^{(2)}(\boldsymbol{x},\eta)+\cdots+\delta_{\rm g}^{(n)}(\boldsymbol{x},\eta), \tag{12.85}$$

where, as we now know, $\delta_{\rm g}^{(1)}=b_1\delta^{(1)}$. A crucial difference from the case of the matter density field is that we have to identify which bias terms need to be included in $\delta_{\rm g}^{(n)}$, i.e. at a given order in perturbation theory, in order to describe a general galaxy density field. As described in detail in Sect. 2 of Desjacques et al. (2018), there is a rigorous theory behind this, which we will not go into here. At second order, in $\delta_{\rm g}^{(2)}$, there are two bias terms: the $b_2$ term we encountered above, and another term involving the tidal field squared, proportional to $b_{K^2}(\partial_i\partial_j\Psi)(\partial^i\partial^j\Psi)$ (see Exercise 12.12). The tidal field did not appear in the thresholding toy model, since we assumed that the halo density only depends on the local value of the matter density. In reality, halo and galaxy formation are influenced by large-scale tidal fields, so we have to include them in the bias relation.

Just as we did for the matter density field based on Eq. (12.29), we can use Eq. (12.85) to expand the galaxy density field in Fourier space by defining kernels $F_{{\rm g},n}$ in analogy to the $F_n$ for matter, Eq. (12.40):

$$\delta_{\rm g}^{(n)}(\boldsymbol{k},\eta)=D_+^n(\eta)\left[\prod_{i=1}^n\int\frac{d^3k_i}{(2\pi)^3}\right](2\pi)^3\delta_{\rm D}^{(3)}\left(\boldsymbol{k}-\sum_{i=1}^n\boldsymbol{k}_i\right)\\ \times F_{{\rm g},n}(\boldsymbol{k}_1,\cdots,\boldsymbol{k}_n;\eta)\delta_0(\boldsymbol{k}_1)\cdots\delta_0(\boldsymbol{k}_n). \tag{12.86}$$

For example, you can show in Exercise 12.12 that the second-order kernel is given by

$$F_{{\rm g},2}(\boldsymbol{k}_1,\boldsymbol{k}_2;\eta)=b_1(\eta)F_2(\boldsymbol{k}_1,\boldsymbol{k}_2)+\frac{1}{2}b_2(\eta)+b_{K^2}(\eta)\left[\frac{(\boldsymbol{k}_1\cdot\boldsymbol{k}_2)^2}{k_1^2k_2^2}-\frac{1}{3}\right]. \tag{12.87}$$

Since the bias parameter $b_1$ multiplies the matter density field, which itself has nonlinear contributions, we obtain a term $b_1F_2$ in $F_{{\rm g},2}$. Further, $b_2$ appears as expected, in addition to the tidal bias parameter $b_{K^2}$. For halos of a given mass, $b_1$ and $b_2$ can be obtained from the peak-background split described above. For observed galaxies, these coefficients need to be determined from the data, by measuring their statistics such as the galaxy power spectrum.

Based on Eqs. (12.85)–(12.86), all calculational techniques, including diagrams, that we developed for matter in Sect. 12.2 carry over to galaxies. The bispectrum of galaxies, for example, can be derived in analogy to Eq. (12.51):

$$\left\langle\delta_{\rm g}(\boldsymbol{k}_1,\eta)\delta_{\rm g}(\boldsymbol{k}_2,\eta)\delta_{\rm g}(\boldsymbol{k}_3,\eta)\right\rangle=(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3)\\ \times\left[2F_{{\rm g},2}(\boldsymbol{k}_1,\boldsymbol{k}_2;\eta)P_{\rm L}(k_1,\eta)P_{\rm L}(k_2,\eta)+B_N(k_1,\eta)+2\ \text{perm.}\right], \tag{12.88}$$

where

$$B_N(k,\eta)=\frac{1}{3}B_{N0}(\eta)+b_1(\eta)P_{N,\delta}(\eta)P_{\rm L}(k,\eta) \tag{12.89}$$

is the noise contribution to the galaxy bispectrum, analogous to $P_N$ in Eq. (11.23). In the galaxy bispectrum, there are two constant noise amplitudes, $B_{N0}$ and $P_{N,\delta}$. Notice the appearance of the power spectrum with the latter; in fact, $P_{N,\delta}$ can be interpreted as a noise term in the linear bias $b_1$.

So far, we discussed galaxy statistics without any observational effects. In particular, we ignored redshift-space distortions (RSD; Sect. 11.1.2). In order to incorporate RSD, we need to justify one more assumption made in our derivation of the observed galaxy clustering in Ch. 11: the fact that large-scale galaxy velocities are unbiased. For matter, we found the velocity by solving the geodesic equation, which describes the motion of massive particles freely falling in a gravitational field $\nabla\Psi$. Now, the equivalence principle of general relativity ensures that *any* free-falling massive particle follows the same geodesic and thus attains the same velocity. This is the fundamental reason why galaxy velocities are unbiased. At some point we do expect galaxy velocities to depart from those of dark matter particles: first, there are non-gravitational interactions at play leading to pressure forces on the gas that makes up galaxies. Second, even the center-of-mass velocities of halos are not the same as those of dark matter particles, since they are given by an average over a large number of particles. Both of these effects, however, are restricted to small scales: we already saw in Eq. (12.53) that the effect of pressure on matter scales as $k^2\delta_\mathrm{m}(\boldsymbol{k})$, and the effect on velocities similarly scales as $k^2\boldsymbol{u}_\mathrm{m}(\boldsymbol{k})$; the same applies to the effect of averaging. Then, RSD can be incorporated beyond linear order by expanding the relation between the intrinsic and redshift-space galaxy density, Eq. (11.6), in perturbations (see Sect. 9 of Desjacques et al., 2018).

This concludes our discussion of how galaxy clustering can be treated within perturbation theory. The perturbative approach has the advantage of allowing for an inclusion of all possible effects that enter in the galaxy–matter relation at a given order in perturbation theory, and is thus very robust. On the other hand, we can use this approach only on those scales where perturbation theory is valid ($k < k_\mathrm{NL}(z)$). There are alternative, empirical simulation-based approaches that do not rely on perturbation theory:

- The ***halo occupation distribution*** (HOD) approach assumes a probability distribution $P(N_\mathrm{g}|M_\mathrm{h})$ for finding $N_\mathrm{g}$ galaxies in a halo of mass $M_\mathrm{h}$ (along with a distribution of positions and velocities of these galaxies within halos). This is then applied to halo catalogs identified in N-body simulations, to obtain a corresponding galaxy catalog. There are typically several free parameters in the distribution $P(N_\mathrm{g}|M_\mathrm{h})$, which can be determined by measuring the statistics of galaxies such as the power spectrum on this catalog and requiring them to match observations.
- The ***abundance matching*** technique instead is based on high-resolution (but gravity-only) N-body simulations which also resolve the substructure within halos. One then assumes that the most luminous or massive galaxies reside in the most massive bound substructures of halos. For example, massive elliptical galaxies are usually assigned to the main, central substructure of halos. Apart from the ambiguity of mass definition (see Sect. 12.4.1), which is even more acute for substructure, this approach has fewer free parameters but still appears to describe galaxy clustering well empirically.

The downside of empirical approaches such as HOD and abundance matching is that they are built on strongly simplified assumptions about the connection between galaxies and halos. It is difficult to rigorously assess the accuracy of these underlying assumptions, since their deficiencies could be partially absorbed by the free parameters involved, or, even worse, by a shift in cosmological parameters. On the other hand, we have a more rigorous control on systematic errors in perturbation theory, since we can estimate how large the next higher-order contribution is (again, this only works on sufficiently large scales where perturbation theory is applicable). Clearly then, it is important to try to infer cosmology using all of these approaches in order to have independent cross-checks.

## 12.7 The halo model

The halo model uses dark matter halos as building blocks of structure to construct a useful empirical model for the nonlinear matter density (see Cooray and Sheth, 2002 for a review). The basic assumption is that *each dark matter particle belongs to one and only one halo*. Using this assumption, we can combine three ingredients—the density profile, mass function, and clustering of halos—to model the statistics of the nonlinear matter density. We will describe this approach briefly here, and relegate more detailed derivations to exercises.

The fundamental assumption of the halo model is that the matter density field consists of a superposition of halos at locations $\boldsymbol{x}_i$ with masses $M_i$, so that it can be written as

$$\rho_{\rm m}^{\rm HM}(\boldsymbol{x}) = \sum_{\text{halos } i} \rho_{\rm h}(|\boldsymbol{x} - \boldsymbol{x}_i|, M_i), \tag{12.90}$$

where $\rho_{\rm h}(x, M)$ is the halo density profile introduced in Sect. 12.4.1 which is assumed to be spherically symmetric for simplicity, and depends only on the mass $M$. Throughout, we will drop the time arguments, since they do not play a role in the derivation. Let us first turn the sum over halos in Eq. (12.90) into an integral over the number density of halos, which in turn is an integral over the local mass function: $n_{\rm h}(\boldsymbol{x}) = \int d\ln M dn(\boldsymbol{x})/d\ln M$. Eq. (12.90) becomes

$$\rho_{\rm m}^{\rm HM}(\boldsymbol{x}) = \int d^3x' \int d\ln M \, \frac{dn(\boldsymbol{x}')}{d\ln M} \, \rho_{\rm h}(|\boldsymbol{x} - \boldsymbol{x}'|, M), \tag{12.91}$$

where $\boldsymbol{x}'$ is the center-of-mass position of the halo that contributes to $\rho_{\rm m}^{\rm HM}(\boldsymbol{x})$. Eq. (12.91) is a spatial convolution, reminiscent of the smoothing operation in Eq. (12.2); indeed, we are spreading the mass $M$ contained in the halo over a region defined by its density profile. So let us define the *normalized profile* $y(x, M) \equiv \rho_{\rm h}(x, M)/M$, which by definition of the profile and total halo mass $M$ obeys

$$\int d^3x \, y(x, M) = 1, \tag{12.92}$$

as usually required for a smoothing filter. This yields

$$\rho_{\rm m}^{\rm HM}(\boldsymbol{x}) = \int d^3x' \int d\ln M \, \frac{dn(\boldsymbol{x}')}{d\ln M} M \, y(|\boldsymbol{x}-\boldsymbol{x}'|, M). \tag{12.93}$$

Now we separate both $\rho_{\rm m}$ and $dn/d\ln M$ into their respective homogeneous parts and perturbations, writing

$$\begin{aligned} \rho_{\rm m}^{\rm HM}(\boldsymbol{x}) &= [1+\delta_{\rm m}^{\rm HM}(\boldsymbol{x})]\rho_{\rm m}, \\ \frac{dn(\boldsymbol{x})}{d\ln M} &= [1+\delta_{\rm h}(\boldsymbol{x}, M)]\frac{dn}{d\ln M}. \end{aligned} \tag{12.94}$$

We obtain

$$1+\delta_{\rm m}^{\rm HM}(\boldsymbol{x}) = \int d\ln M \, \frac{M}{\rho_{\rm m}} \frac{dn}{d\ln M} \int d^3x' [1+\delta_{\rm h}(\boldsymbol{x}', M)] y(|\boldsymbol{x}-\boldsymbol{x}'|, M). \tag{12.95}$$

We can now use the normalization of the halo mass function. Recall that by the basic assumption of the halo model, all matter particles are contained in halos. This means that the integral over the halo mass function, weighted by mass, must yield the mean matter density:

$$\int d\ln M \, M \frac{dn}{d\ln M} = \rho_{\rm m}. \tag{12.96}$$

This corresponds to the condition Eq. (12.71) for the collapsed fraction (and places an additional constraint on the function $f(\nu)$ in Eq. (12.73)). Using this result, Eq. (12.95) becomes

$$\delta_{\rm m}^{\rm HM}(\boldsymbol{x}) = \int d\ln M \, \frac{M}{\rho_{\rm m}} \frac{dn}{d\ln M} \int d^3x' \delta_{\rm h}(\boldsymbol{x}', M) y(|\boldsymbol{x}-\boldsymbol{x}'|, M). \tag{12.97}$$

This equation shows that we can compute the nonlinear matter power spectrum in the halo model based on the mass function of halos ($dn/d\ln M$), the halo profile ($y(x, M)$), and the clustering of halos (via the power spectrum of $\delta_{\rm h}$). You can work out the details in Exercise 12.13.

It turns out that the matter power spectrum in the halo model naturally breaks down into two components, $P^{\rm HM}(k) = P_{\rm 2h}(k) + P_{\rm 1h}(k)$: the first, "two-halo term," is due to the large-scale clustering of halos themselves, i.e. involves two mass elements in different halos (see Fig. 12.11). The second, "one-halo term" $P_{\rm 1h}$ is essentially the halo shot noise convolved with their profile. This term would even be present if halos were not clustered at all; it involves only mass elements within a single halo. We can guess from Fig. 12.11 that it will be the dominant term on small scales.

A numerical calculation of the resulting power spectrum is shown in Fig. 12.12. On large scales, it agrees with the linear power spectrum, as required. On those scales, the 2-halo term dominates. On small scales, the 1-halo term dominates. In this regime, we are probing the interior of halos, so that the power spectrum is given by a combination of halo

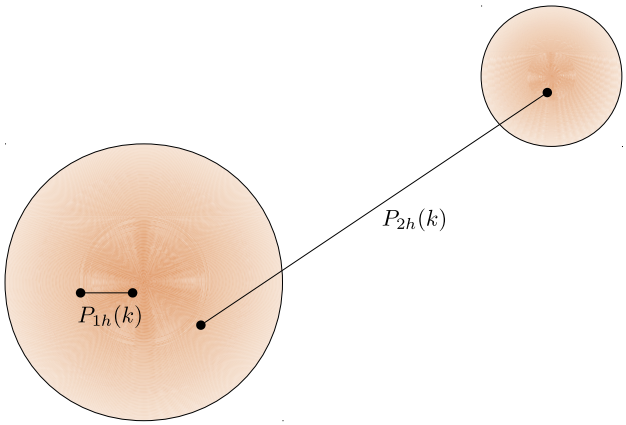



**FIGURE 12.11** Illustration of the contributions to the halo-model power spectrum. All matter is in halos of various mass (shaded circles). On large scales, the power spectrum is dominated by contributions involving two halos ($P_{2\mathrm{h}}(k)$). On small scales, the power spectrum is dominated by contributions from within a single halo ($P_{1\mathrm{h}}(k)$).

profile shapes and the mass function. The bottom panel of Fig. 12.12 shows the comparison of this simple calculation with the nonlinear matter power spectrum measured in full N-body simulations. Given its simplicity, the halo model does remarkably well, predicting the power spectrum to within 25% over a wide range of scales, with the most significant departure being an underprediction of the power spectrum on small scales (which can be improved by tuning the assumed concentration–mass relation of halos). For precision applications, this is not sufficient, but the wide range of scales that are described approximately does illustrate that the halo model can be a useful phenomenological framework.

Using the same assumptions we have made for the matter power spectrum, other statistics, such as the cross-correlation between matter and halos, or the bispectrum of matter can likewise be computed.

## 12.8 Summary

For the first time in this book, we went beyond small, linear perturbations in cosmology in this chapter, studying the evolution of nonlinear structure. Essentially all observables of large-scale structure, including galaxy clustering, cluster abundance, and weak lensing (the topic of Ch. 13) are significantly impacted by nonlinear evolution. Understanding nonlinear structure formation has thus become an integral part of modern cosmology. Fortunately, this problem remains tractable: first, the metric is still well described by FLRW with small perturbations; only the matter density and velocity have to be treated nonlinearly. Second, the complexities of baryonic physics (pressure, cooling, star formation, and so on) only significantly impact the matter distribution on very small scales ($k \gtrsim 1\,h\,\mathrm{Mpc}^{-1}$); so, as long as we content ourselves with larger, yet still nonlinear scales, we can treat baryons and dark matter as a single component of cold collisionless matter.

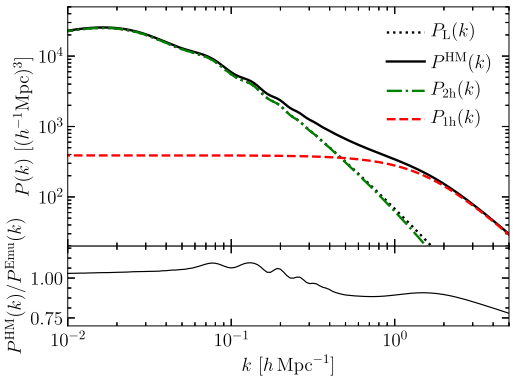



**FIGURE 12.12** *Upper panel:* The nonlinear matter power spectrum in the halo model at $z = 0$ (black solid). Here, the prescription for $f(\nu)$ shown in Fig. 12.9 was used along with the corresponding bias derived through Eq. (12.80), and a truncated NFW profile (see Exercise 12.13 for details). We also show the 1-halo ($P_{\rm 1h}$) and 2-halo ($P_{\rm 2h}$) contributions, as well as the linear power spectrum. Notice the small constant contribution to which $P_{\rm 1h}(k)$ asymptotes on large scales, which is unphysical but numerically unimportant. *Lower panel:* ratio of the halo-model power spectrum to that measured in N-body simulations, interpolated to the fiducial cosmology using the CosmicEmu code (Heitmann et al., 2014). The halo model predicts the matter power spectrum to within 25% out to very small scales, while it does much better on large scales.

The fundamental set of equations governing the nonlinear growth of matter is the collisionless, non-relativistic Boltzmann (Vlasov) equation coupled to gravity via the Poisson equation:

$$\frac{\partial f_{\rm m}}{\partial t} + \frac{\partial f_{\rm m}}{\partial x^j}\frac{p^j}{ma} - \frac{\partial f_{\rm m}}{\partial p^j}\left[Hp^j + \frac{m}{a}\frac{\partial \Psi}{\partial x^j}\right] = 0,$$
$$\nabla^2\Psi = 4\pi G a^2\left[\int \frac{d^3p}{(2\pi)^3} f_{\rm m}(\boldsymbol{x}, \boldsymbol{p}, t) - \rho_{\rm m}(t)\right]. \tag{12.98}$$

This conceptually simple, but mathematically complex set of equations leads to a rich set of predictions. We only scratched the surface in this chapter, and refer those who want to delve deeper to the excellent comprehensive review of Bernardeau et al. (2002), which covers both numerical and perturbation-theory approaches.

One powerful approach to the Vlasov–Poisson system is to solve it numerically via N-body simulations, which discretize the phase-space volume occupied by matter (this volume takes the form of a thin sheet due to the cold nature of dark matter). Thus, the particles in these simulations are not actually dark matter particles, but stand for chunks of phase space. The output of N-body simulations allows us to identify massive bound structures, dark matter halos, which host galaxies and clusters of galaxies. The mass $M_\Delta$ of halos is defined as that within a comoving radius $R_\Delta$ enclosing a mean interior density that is $\Delta$ times the mean matter density:

$$M_\Delta = M(< R_\Delta) = \frac{4\pi}{3} R_\Delta^3 \rho_{\rm m}(t_0)\Delta. \tag{12.99}$$

Another important scale is the *Lagrangian radius* $R_L$ which encloses the mass of the halo at the mean matter density:

$$R_L(M) = 1.40\,h^{-1}\,\mathrm{Mpc}\left(\frac{M}{10^{12}\,h^{-1}\,M_\odot}\right)^{1/3}\left(\frac{\Omega_\mathrm{m}}{0.31}\right)^{-1/3}. \tag{12.100}$$

The Lagrangian radius gives the comoving size of the region from which matter assembled to form the halo, and plays a major role in semi-analytic approaches to describing halos. For example, the number density of halos per logarithmic mass interval, the halo mass function, is usually parametrized as

$$\frac{dn(M,z)}{d\ln M} = \frac{\rho_\mathrm{m}(t_0)}{M} f\left(\frac{\delta_\mathrm{cr}}{\sigma(M,z)}\right)\left|\frac{d\ln\sigma(M,z)}{d\ln M}\right|, \tag{12.101}$$

where $\sigma(M,z) \equiv \sigma(R_L[M],z)$ is the variance of the linear density field filtered with a real-space tophat filter on the scale of the Lagrangian radius $R_L(M)$, and $f(\nu)$ is a fitting function. By tuning a single function $f(\nu)$, Eq. (12.101) can be made to describe the halo mass function for a range of redshifts and cosmologies to $\sim 5\%$ accuracy (Fig. 12.9).

The relation between halos and observable objects is particularly clear for clusters, which are the most massive and rare bound structures in the universe. A measurement of the halo abundance, together with a relation between cluster observables such as X-ray temperature or the SZ distortion parameter $y$, and halo mass then allows for cosmology constraints from clusters. Such a relation can in particular be obtained from gravitational lensing, the topic of the next chapter.

Halos and their approximately universal density profiles also allow us to build an empirical picture for the nonlinear matter density field, known as the halo model, by positing that each dark matter particle belongs to one (and only one) halo. We only briefly introduced the idea, and we refer the interested reader to the review by Cooray and Sheth (2002), as well as Exercise 12.13. Toy models such as this, as well as the spherical collapse picture of halo formation, are not very accurate, but give physical intuition and allow us to explore how nonlinear structure reacts to new physics (such as massive neutrinos or modifications to general relativity).

While gravity-only N-body simulations are a routine affair by now, they do not produce galaxies, or any sources of light for that matter, since they do not simulate electromagnetic and other interactions of baryons. Simulating baryons realistically necessarily requires recipes for modeling unresolved scales, and how to do this accurately is still the topic of much ongoing research. Readers interested in a more in-depth treatment of these and many other aspects of nonlinear structure and galaxy formation should consult the book by Mo et al. (2010).

A key takeaway message then is that, since we do need to understand the behavior of galaxies in order to interpret measurements of galaxy clustering, we cannot rely exclusively on simulations. An alternative approach is to stay close to the analytic route we have pursued in most of the book so far, by performing a perturbation-theory expansion of the nonlinear matter density field:

$$\delta_{\rm m}(\boldsymbol{x},\eta) = \delta^{(1)}(\boldsymbol{x},\eta) + \delta^{(2)}(\boldsymbol{x},\eta) + \cdots + \delta^{(n)}(\boldsymbol{x},\eta),$$
$$\theta_{\rm m}(\boldsymbol{x},\eta) = \theta^{(1)}(\boldsymbol{x},\eta) + \theta^{(2)}(\boldsymbol{x},\eta) + \cdots + \theta^{(n)}(\boldsymbol{x},\eta). \tag{12.102}$$

The equations for $\delta_{\rm m}^{(n)}$ and $\theta_{\rm m}^{(n)}$ are obtained by taking moments of the Vlasov equation, and by truncating the hierarchy at the second moment. The result is a set of equations describing an effective, pressureless fluid:

$$\delta_{\rm m}{}' + \theta_{\rm m} = -\delta_{\rm m}\theta_{\rm m} - u_{\rm m}^j \frac{\partial}{\partial x^j}\delta_{\rm m},$$
$$\theta_{\rm m}{}' + aH\theta_{\rm m} + \nabla^2\Psi = -u_{\rm m}^j \frac{\partial}{\partial x^j}\theta_{\rm m} - (\partial_i u_{\rm m}^j)(\partial_j u_{\rm m}^i),$$
$$\nabla^2\Psi = \frac{3}{2}\Omega_{\rm m}(\eta)(aH)^2\delta_{\rm m}. \tag{12.103}$$

By substituting lower-order solutions into the nonlinear source terms on the right-hand side, we can calculate the matter evolution order by order. We used this to compute the nonlinear correction to the matter power spectrum in Eq. (12.48),

$$P(k,\eta) = P_{\rm L}(k,\eta) + P^{\rm NLO}(k,\eta). \tag{12.104}$$

The major downside of the perturbative approach is that its range of validity is restricted to large scales, i.e. scales where the second term in Eq. (12.104) is smaller than the first (wavenumbers $k \lesssim 0.2\,h\,{\rm Mpc}^{-1}$ at $z = 0$, but extending to increasingly larger wavenumbers at higher redshifts). The major advantage of perturbation theory is that we can robustly calculate the clustering of matter and galaxies with minimal assumptions about small-scale baryonic effects, which are then captured by bias coefficients. As we discussed, this robustness is especially important for galaxies, for which we need to rely on a bias relation in order to be able to infer cosmology. In Sect. 12.6 we used this to justify the assumptions made about galaxy clustering in the previous chapter, and to extend them to higher orders in perturbation theory.

## Exercises

**12.1** Show that the stress tensor $\sigma_{\rm m}^{ij}$ [Eq. (12.17)] vanishes for a "cold" distribution function of the form Eq. (12.9).

**12.2** Use Eq. (12.23) to derive an equation for the vorticity $\boldsymbol{\omega} = \boldsymbol{\nabla} \times \boldsymbol{u}_{\rm m}$ of the matter velocity. Show that no vorticity is generated if it is absent in the initial conditions. How does an initial vorticity evolve in time at linear order?

**12.3** Fill in the missing steps of the transformation of the Euler–Poisson system into Fourier space, Eq. (12.31).

**12.4** Use the equation for the linear growth factor Eq. (8.75) to prove Eq. (12.32). Note that this relation holds for any smooth dark energy model. Next, use this to transform Eq. (12.31) into Eqs. (12.33)–(12.34).

**12.5** Use the solution in Eq. (12.40) to show that the NLO contribution in Eq. (12.42) is given by Eq. (12.48). Derive and use the relations

$$F_2(\boldsymbol{k}, -\boldsymbol{k}) = 0,$$
$$F_n(\boldsymbol{k}_1, \cdots, \boldsymbol{k}_n) = F_n(-\boldsymbol{k}_1, \cdots, -\boldsymbol{k}_n). \tag{12.105}$$

Evaluate the terms numerically. For $P^{(13)}$, the expression of the kernel given in Makino et al. (1992) is useful. For $P^{(22)}$, care needs to be taken when $\boldsymbol{k} - \boldsymbol{p}$ becomes close to zero. Bertschinger and Jain (1994) provide a neat decomposition of the integral which is numerically robust.

**12.6** Derive the leading contribution to the matter bispectrum, Eq. (12.51). How does this look in the diagram form of Fig. 12.3?

**12.7** In Sect. 12.2, we developed perturbation theory based on the density field. An alternative, *Lagrangian* approach is based on the equations of motion for N-body "particles," Eq. (12.57). In this exercise, you will derive the lowest-order result, known as *Zel'dovich approximation*. The solution to Eq. (12.57) is a particle trajectory $\boldsymbol{x}(\eta)$. We write this as

$$\boldsymbol{x}(\eta) = \boldsymbol{q} + \boldsymbol{s}(\boldsymbol{q}, \eta), \tag{12.106}$$

where $\boldsymbol{q}$ is the initial position at $\eta = 0$, when all perturbations were negligible. Hence $\boldsymbol{s}(\boldsymbol{q}, 0) = 0$. Rewrite Eq. (12.57) as an equation for $\boldsymbol{s}$. Now expand to linear order in $\boldsymbol{s}$. Solve the equation by using the solution of the Poisson equation for $\Psi$ at linear order. Your result should relate $\boldsymbol{s}^{(1)}(\boldsymbol{k}, \eta)$ to $\delta^{(1)}(\boldsymbol{k}, \eta)$. This result can be used to obtain the initial small displacements of particles to start an N-body simulation. We also need their initial momenta $p_c^i$. Derive these in terms of the displacement as well.

**12.8** Derive an expression for the enclosed mass $M(< r)$ for the NFW profile Eq. (12.62). Replace $r_s$ with the concentration $c_\Delta$. Use this to derive $R_\Delta$ for a given mass $M_\Delta$ and concentration, and solve for $\rho_s$. You now have a reasonably accurate expression for the density profile of a halo of mass $M_\Delta$ and concentration $c_\Delta$. Plot the profile for a halo of mass $M_{200} = 10^{12} M_\odot$ ($\Delta = 200$), and for concentrations $c_{200} \in \{4, 8, 16\}$. That is, make the plot for Sect. 12.4.1 that we were too lazy to create!

**12.9** Derive the spherical collapse threshold $\delta_{\rm cr}$ and the virial overdensity $\Delta_{\rm vir}$ by solving Eq. (12.67) without considering $\Lambda$. Follow these steps:

**(a)** Show that Eq. (12.67) can be rewritten as

$$\frac{\ddot{r}}{r} = -\frac{4\pi G}{3} \bar{\rho}_i [1 + \delta_i] \left(\frac{r_i}{r}\right)^3 \tag{12.107}$$

where $r_i$, $\bar{\rho}_i$ are, respectively, the radius of the spherical region and the background matter density at the initial time, and $\delta_i$ is the initial overdensity.

**(b)** Show that, when the initial expansion rate is given by $\dot{r}_i = H_i r_i (1 - \delta_i/3)$, the maximum radius $r_{\rm ta}$ (the turn-around radius) that the spherical region reaches is given by

$$r_{\rm ta} = r_{\rm ta} = \frac{3}{5}\left(\frac{1+\delta_i}{\delta_i}\right) r_i\,. \tag{12.108}$$

**(c)** Show that the parametric solution (cycloid) of Eq. (12.68) is a solution of Eq. (12.107). What is $t_{\rm ta}$ in terms of the initial conditions, $r_i$, $\delta_i$, $\bar{\rho}_i$?

**(d)** Find the expression for the nonlinear overdensity $\delta(\theta)$. What is the nonlinear density contrast at the time of turn around? Plot $\delta(t)$ as a function of $\delta^{(1)}(t)$, the initial overdensity evolved forward using the linear growth factor. Derive the expansion of $\delta(\delta^{(1)})$ to third order in $\delta^{(1)}$.

**(e)** Assume that, by some magic (which we call violent relaxation), the object virializes. Find the virial radius in terms of the turn around radius. Using that, give the density contrast $\Delta_{\rm vir} \equiv 1 + \delta(t_{\rm vir})$ expected after virialization. Assuming that collapse is completed at $\theta = 2\pi$ [that is, $t_{\rm vir} = t(\theta = 2\pi)$], what is the value of $\delta^{(1)}(t)$ at collapse? This is the collapse threshold $\delta_{\rm cr}$.

**12.10** Derive the correlation function of thresholded regions Eq. (12.82) in the linear density field.

**(a)** Define the scaled density field $\nu(\boldsymbol{x}) \equiv \delta_R^{(1)}(\boldsymbol{x})/\sigma(R)$ (note that this is a *field*, and not to be confused with the parameter $\nu = \delta_{\rm cr}/\sigma(R)$, which we shall indicate with $\nu_{\rm cr}$ in this exercise). Show that $\nu(\boldsymbol{x})$ at an arbitrary fixed location follows the normal distribution

$$p(\nu) = \frac{1}{\sqrt{2\pi}} e^{-\nu^2/2}, \tag{12.109}$$

and that the joint distribution of $\nu_1$, $\nu_2$, where $\nu_i \equiv \nu(\boldsymbol{x}_i)$, is a bivariate Gaussian

$$p(\nu(\boldsymbol{x}_1), \nu(\boldsymbol{x}_2)) = \frac{1}{2\pi\sqrt{1-\xi_{12}^2/\sigma^4(R)}} \exp\left[-\frac{1}{2}(\nu_1,\nu_2)^{\top} \mathrm{C}^{-1} (\nu_1,\nu_2)\right] \tag{12.110}$$

$$\text{where} \quad \mathrm{C} = \begin{pmatrix} 1 & \xi_{12}/\sigma^2(R) \\ \xi_{12}/\sigma^2(R) & 1 \end{pmatrix} \tag{12.111}$$

and $\xi_{12} = \xi_R^{(1)}(|\boldsymbol{x}_1 - \boldsymbol{x}_2|) = \left\langle \delta_R^{(1)}(\boldsymbol{x}_1)\delta_R^{(1)}(\boldsymbol{x}_2)\right\rangle$ is the correlation function of the linear, smoothed density field.

**(b)** Using this result, show that the one-point probability (or volume fraction) becomes

$$p(\delta_R^{(1)} > \delta_{\rm cr}) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\nu_{\rm cr}}{\sqrt{2}}\right). \tag{12.112}$$

Here, $\nu_{\rm cr} \equiv \delta_{\rm cr}/\sigma(R)$ and the complementary error function is defined in Eq. (C.31). Obtain the corresponding expression for the joint probability

$$p\left(\delta_R^{(1)}(\boldsymbol{x}_1) > \delta_{\rm cr}, \delta_R^{(1)}(\boldsymbol{x}_2) > \delta_{\rm cr}\right), \tag{12.113}$$

and for $\xi_{\rm thr}(r)$ from Eq. (12.82). Notice that one of the two integrals can be done analytically.

**(c)** Using the fact that the matter correlation function goes to zero at large $r$, expand your result in the small quantity $\xi(r)$. Show that the first two terms can be written as Eq. (12.83), and derive the expression for $b_1$ and $b_2$, as well as their limiting form for very rare halos, $\nu_{\rm cr} \gg 1$.

**12.11** Continue the expansion in Eq. (12.80) to second order in $\delta_\ell$. The second-order bias is defined by

$$\delta_{{\rm h},\ell} = b_1\delta_\ell + \frac{1}{2}b_2\delta_\ell^2 + \cdots . \tag{12.114}$$

What is the expression for $b_2$ in terms of $\sigma(M,z)$ and $f(\nu)$? Derive $b_2(\nu)$ for the Press–Schechter mass function Eq. (12.73).

**12.12** Derive the second-order perturbation theory kernel for the galaxy density $\delta_{\rm g}^{(2)}$.

**(a)** Define the scaled tidal field through

$$K_{ij}(\boldsymbol{x},\eta) = \frac{1}{4\pi G a^2(\eta)}\left[\partial_i\partial_j - \frac{1}{3}\delta_{ij}\nabla^2\right]\Psi(\boldsymbol{x},\eta). \tag{12.115}$$

Use this definition to relate $K_{ij}$ is to the matter density in real and Fourier space.

**(b)** Begin with the real-space expression of the second-order galaxy density,

$$\delta_{\rm g}^{(2)}(\boldsymbol{x},\eta) = b_1\delta^{(2)} + \frac{1}{2}b_2(\delta^{(1)})^2 + b_{K^2}K_{ij}^{(1)}K^{(1)ij}, \tag{12.116}$$

where on the right-hand side all fields are evaluated at $(\boldsymbol{x},\eta)$, and the bias parameters $b_1$, $b_2$, $b_{K^2}$ are defined at $\eta$. Why does the tidal field only appear at second order and in this particular combination? Now pull out the time dependence contained in the growth factors, and Fourier transform Eq. (12.116) to arrive at Eq. (12.87).

**12.13** Compute the matter power spectrum in the halo model based on Eq. (12.97).

**(a)** Take the Fourier transform of Eq. (12.97), and express the power spectrum of $\delta_{\rm m}^{\rm HM}(\boldsymbol{k})$ in terms of the power spectrum of the halo overdensity $\delta_{\rm h}(\boldsymbol{k},M)$, the halo mass function, and the Fourier-transform of the halo profile $y(k,M)$.

**(b)** Assume a linear bias relation and constant noise for halos:

$$\left\langle\delta_{\rm h}(\boldsymbol{k},M)\delta_{\rm h}(\boldsymbol{k}',M')\right\rangle = (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}')\left[b_1(M)b_1(M')P_{\rm L}(k) + P_N(M,M')\right] \tag{12.117}$$

$$\text{where}\quad P_N(M,M') = \frac{1}{dn/d\ln M}\delta_{\rm D}^{(1)}(\ln M - \ln M').$$

Here we have assumed that the noise for different halo masses is independent. Use this to simplify the expression for the halo-model matter power spectrum.

**(c)** Derive the Fourier-transform of the profile $y(k,M)$ by assuming an NFW profile Eq. (12.62) with concentration $c(M)$ that is truncated at $R_{200}$.

**(d)** Evaluate $P^{\rm HM}(k)$ numerically, and plot the result together with the linear power spectrum.

# 13

# Probes of structure: lensing

The methods of probing large-scale structure we encountered so far—galaxy clustering, SZ effect, and cluster number counts—are powerful probes of cosmology but share a common deficiency. They are measures of the distribution of galaxies, or more generally baryonic matter, not the distribution of mass. It is much easier to make accurate predictions about the latter. As we have seen in Ch. 11 and Ch. 12, the connection between galaxies and matter (bias) can be rigorously trusted only on large scales and at the price of introducing additional free coefficients; the mass-observable relation for galaxy clusters is likewise very difficult to predict theoretically.

A very exciting experimental approach that probes the entire mass distribution is introduced in this chapter. We will see that the inhomogeneities induce distortions in the observed shapes of distant galaxies due to *gravitational lensing*. They also induce small distortions in the CMB which can be detected as they modify the anisotropies from their primordial form. Further, the statistics of these lensing distortions are directly related to the matter power spectrum; specifically, the power spectrum of nonlinear matter we have discussed in detail in Ch. 12, since lensing probes structures down to small scales.

We begin with an overview and introduction to lensing in Sects. 13.1–13.2, before turning to CMB lensing in Sect. 13.3. Gravitational lensing measured through galaxy shapes (*shear*), which we describe next, is based on the same formalism as polarization. For this reason, you should go through Sect. 10.1 before reading Sect. 13.4 and following. The reason is that both effects can be quantified with a two-by-two symmetric matrix on the sky: in Ch. 10, we dealt with the polarization tensor with its $Q$ and $U$ components. In lensing, the analogue is the second-moment tensor of galaxy images. Section 13.5 then combines the results of Ch. 10 and Ch. 11 to derive the weak lensing power spectra and correlation functions.

## 13.1 Overview

The gravitational effect of inhomogeneities in the universe distorts the paths traveled by light from distant sources to us. We already encountered this effect, known as *gravitational lensing* (or *lensing* for short), when studying photon geodesics in Sect. 3.3.2. Lensing is so promising because light paths respond to *mass*; more precisely, lensing probes all clustering stress-energy components in the universe via the spacetime perturbations $\Phi$, $\Psi$. If we can measure these distortions, then, we infer something about the distribution of mass in the universe. This makes lensing highly complementary to the other probes of structure discussed in Ch. 11.

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00019-X
Copyright © 2021 Elsevier Inc. All rights reserved.

The idea that gravitational fields might lens distant images is at least as old as general relativity. Indeed, even before Einstein finalized general relativity, he understood the importance of measuring this distortion. Early notebooks of his contain calculations of the magnification of images and of the possibility of a double image of a single source (Renn et al., 1997). And it was detection of gravitational lensing that led to the acceptance of general relativity. In 1919, Eddington led a voyage to the Southern Hemisphere to observe the deflection of starlight during a solar eclipse. The magnitude of this effect (Dyson et al., 1920) was in good agreement with Einstein's new theory.

In 1979, Walsh, Carswell, and Weymann observed a multiply imaged QSO, thereby confirming Einstein's early speculations. Light rays leaving the QSO in different directions are focused on the same point (us) by an intervening galaxy. The number of lensed QSO can be shown to be a probe of the cosmological volume as a function of redshift, and hence the expansion history (Kochanek, 1996).

There are other examples of gravitational lensing that have an impact on cosmology, such as time delays: two light rays emitted from the same source at the same time which we detect from different directions due to lensing typically arrive at different times. We can measure this time delay by studying sources with variable emission. The delay turns out to depend on the Hubble constant, so astronomers have used this technique to measure $H_0$ (e.g., Wong et al., 2019). Another example is microlensing, wherein a lens moves into the line connecting a source to us. When it does, the image is magnified, so that we observe a characteristic variability in the distant source. Microlensing has been used to constrain the contribution of massive compact halo objects (MACHOs) to the dark matter (e.g., Tisserand et al., 2007). Lensing has also been employed to search for the substructure that we expect, from Sect. 12.3, to form in cold dark matter halos (Vegetti et al., 2012; Hezaveh et al., 2016).

Another spectacular manifestation of gravitational lensing is shown in Fig. 13.1. The large galaxy cluster in the center, whose member galaxies appear yellow (light gray in print version), distorts the images of distant galaxies in the background. Why do the background galaxies appear stretched out as arcs in Fig. 13.1? Consider a galaxy whose light passes close to a massive lens on the way to the observer (left panel in Fig. 13.2). The apparent position of the galaxy is displaced away from the lens, since the rays are bent toward it. The right panel of the figure shows the effect that this has on the lensed image of the source galaxy, assuming for simplicity that the lens has cylindrical symmetry around the line of sight. Each pixel in the source is displaced radially outward, distorting the regular, elliptical unlensed image into an arc. Another distorting effect is due to the fact that rays passing closer to the lens are distorted more, so that the arc is narrower in the radial direction than the unlensed image.

In the next section, we will show that lensing conserves the surface brightness of an image. Since the lensed image has a bigger area than the unlensed one, the net effect is an increase in the total flux of the source galaxy image; this is known as *magnification*. Using lensing, it is possible to find and study galaxies behind massive lenses that would otherwise be much too faint to detect.



**FIGURE 13.1** The massive galaxy cluster Abell 1689, as imaged by the Hubble space telescope. Lensed galaxies behind the cluster are visible as strongly distorted arcs. These high-redshift galaxies often appear blue, because a break in the spectrum leads to them dropping out of the red imaging bands. Credit: Author NASA, ESA, the Hubble Heritage Team (STScI/AURA), J. Blakeslee (NRC Herzberg Astrophysics Program, Dominion Astrophysical Observatory), and H. Ford (JHU). CC-SA 4.0.

Another magnificent example of gravitational lensing is shown in Fig. 13.3. The large double cluster in the foreground, the so-called bullet cluster, distorts the shapes of the background galaxies. This leads to a distinctive pattern of elliptical arcs surrounding the two separate dark matter cores of the cluster. The bullet cluster is an ongoing merger between two massive galaxy clusters, and provides striking evidence for the presence of collisionless dark matter. Most of the *baryons* in the two clusters are in the form of hot diffuse gas that is detected in X-rays (Sect. 12.5; shown in pink (light gray in print version) shading in Fig. 13.3). These are clearly in a different location than the bulk of *matter*, detected via gravitational lensing (blue (gray in print version) shading in the figure). This displacement is expected physically: the diffuse gas is collisional, so it formed a strong shock when the two clusters collided. On the other hand, dark matter is collisionless, so the two dark matter halos freely passed through each other, similar to the illustration in Fig. 12.5. The same actually applies to the galaxies contained in each cluster: these also suffered few interactions compared to the diffuse baryons, and so are located where the bulk of the matter is. While we have encountered plenty of extremely robust pieces of ev-

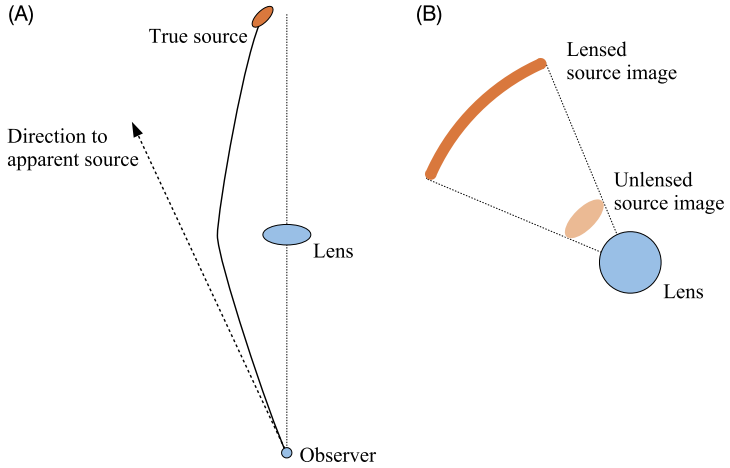



**FIGURE 13.2** *Left panel:* top view of a lensing system. Light emitted from the source (top) reaches the observer at the bottom following a trajectory that is being deflected by the gravitational field of the lens. From the observer's point of view, the apparent position of the source is in the direction of the dashed arrow, further away from the lens than the true source. *Right panel:* Appearance of the system on the sky, assuming that the lens is cylindrically symmetric around the line of sight. Each point of the source is displaced outward radially. This distorts the source image that would be observed in the absence of lensing into a tangential arc.

idence for cold dark matter in this book, the bullet cluster is another compelling piece simply because it offers such a clear visualization.

For cosmology, the most important aspect of gravitational lensing is *weak lensing*, wherein the shapes of distant galaxies are slightly distorted by intervening foreground matter overdensities; that is, it is the same effect as shown in Fig. 13.2, but with much smaller amplitude. One application of this is to use background galaxies to infer the mass of individual galaxy clusters (dating back to at least Tyson et al., 1990). As argued in Sect. 12.5, we can use the abundance of galaxy clusters as a sensitive probe of cosmology, but only if we have a precise calibration of their mass. Weak lensing is able to provide just that. We will briefly describe how this works in Sect. 13.5.3.

In this chapter, we will mostly be interested in weak lensing by the large-scale matter distribution structure in the universe, rather than by a single identifiable lens such as a cluster. Inferring the distribution of the dark matter, i.e. making a *mass map* is only one part of the goal. We are most interested in statistics such as the correlation function or its Fourier transform, the power spectrum. Indeed, we will derive how the power spectrum of the lensing map measures the underlying (nonlinear) matter power spectrum, and how its cross-correlation with galaxy number counts provides valuable constraints on bias.

## 13.2 Photon geodesics

Yet again, the effect of gravitational lensing on the observed photons is an application of the Boltzmann equation. Since we can neglect scattering and absorption in the late

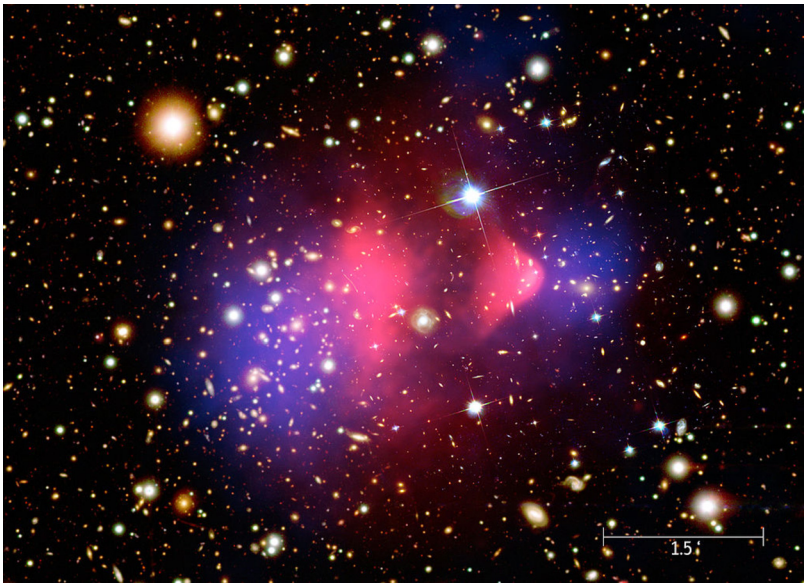



**FIGURE 13.3** The bullet cluster 1E 0657-56. The image in the background was taken in the visible band (Magellan and Hubble Space Telescopes). The mass distribution reconstructed from the distortions of background galaxies is shown as blue (gray in print version) shading, and is displaced from the X-ray emission of the hot gas shown by the pink (light gray in print version) overlay (recorded by the Chandra space telescope). Credit: NASA/CXC/M. Weiss—Chandra X-Ray Observatory: 1E 0657-56.

universe, there is no collision term. Then, the Boltzmann equation simply states that the photon distribution function is conserved: $df(\boldsymbol{x}, \boldsymbol{p}, t)/dt = 0$ (we will not consider polarization in this chapter). Any measurement astronomers and CMB experimentalists make using their detectors can be described as an integral over the specific intensity $I_\nu$, which is defined as the energy incident on a detector per solid angle, per unit area and time, and per unit frequency:

$$dE = I_\nu \, d\Omega \, dA_\perp \, dt \, d\nu, \tag{13.1}$$

where $dA_\perp$ is the detector area orthogonal to the photon flux. All of the information on the radiation field is contained in $f(\boldsymbol{x}, \boldsymbol{p}, t)$, so we should be able to relate this to $I_\nu$. Indeed, we saw a glimpse of this already in our first encounter with $I_\nu$, Eq. (1.9). The precise relation, which you can derive in Exercise 13.1, is

$$I_\nu(\boldsymbol{x}, \hat{\boldsymbol{p}}, t) = 4\pi\nu^3 f(\boldsymbol{x}, p = 2\pi\nu, \hat{\boldsymbol{p}}, t), \tag{13.2}$$

where $\hat{\boldsymbol{p}}$ is the unit momentum vector of the photons that are being detected. The key point is that the specific intensity is directly related to the distribution function of photons, which is *conserved*. And it is the specific intensity that is measured, although we often characterize this measurement using derived quantities (e.g., we extract the temperature of the CMB or the flux of a galaxy image).

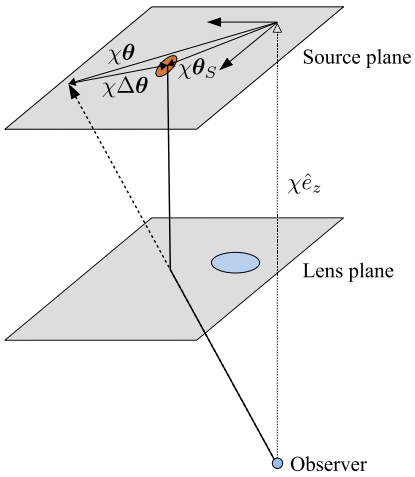



**FIGURE 13.4** Similar lensing setup as shown in the left panel of Fig. 13.2, with our coordinate definitions in the source plane indicated. The origin is given by the intersection of the unit vector $\hat{\boldsymbol{e}}_z$ with the plane of the source. The true source position is at $\chi\boldsymbol{\theta}_S$, while the apparent source position is pointed to by the dashed arrow and defined as $\chi\boldsymbol{\theta}$. The lensing deflection in the source plane is $\chi\Delta\boldsymbol{\theta}$. Here, we have greatly exaggerated the deflection angle.

The conservation of $f$ means that, between emission of the photons and observations, $I_\nu$ changes only through the change in frequency $\nu$, due to the cosmological as well as gravitational redshifts and the Doppler effect. The shift in the frequency only affects the estimated distance to the source, and does not change galaxy shapes. So we will ignore it here. Then, the observed $I_\nu$ at position $\boldsymbol{\theta}$ is the same as would be observed from the direction of the true source $\boldsymbol{\theta}_S$ in the absence of lensing, i.e. in the homogeneous universe (cf. Fig. 13.4):

$$I_{\rm obs}(\boldsymbol{\theta}) = I_{\rm true}(\boldsymbol{\theta}_S). \tag{13.3}$$

The observed intensity at position $\boldsymbol{\theta}$ in the sky then is actually due to a source at position $\boldsymbol{\theta}_S$, but the intensity is unchanged otherwise. Eq. (13.3) is the starting point from which we will derive all our lensing results.

So, we want to solve for the path of a light ray as it leaves a distant source and travels through the inhomogeneous universe. So far in this book, photon paths have always been straight lines in comoving coordinates. Now we allow for a small bending of their trajectories as in Fig. 13.2. Fig. 13.4 shows the geometry and notation. The position of the photon in comoving coordinates at any time is given by $\boldsymbol{x}$, and governed by the geodesic equation we derived in Sect. 3.3.2. Throughout, we will assume a small deflection angle, and work to linear order in this deflection. This is a very good approximation for all applications we will study in this chapter. Then, the $x^3$ ($z$-)component of the position remains equal to the radial distance $\chi$ and the transverse components are equal to $\chi\boldsymbol{\theta}_S$. In other words, the true source location is

$$\boldsymbol{x}_{\rm true} = (\boldsymbol{\theta}_S, 1)\,\chi\,, \tag{13.4}$$

while the apparent location at the same distance is defined through $\boldsymbol{\theta}$,

$$\boldsymbol{x}_{\rm obs} = (\boldsymbol{\theta}, 1)\,\chi\,. \tag{13.5}$$

We have already performed the bulk of the work necessary to obtain $\boldsymbol{\theta}_S$ in terms of $\boldsymbol{\theta}$ in Sect. 3.3.2. We first express the transverse components of the position vector $x^i_\perp$ as an integral over $\chi$ of $dx^i_\perp/d\chi = -dx^i_\perp/d\eta$, since $d\eta = -d\chi$ for light rays (going outward in distance is going backward in time). We have

$$\frac{dx^i_\perp}{d\chi} = -\frac{dx^i_\perp}{d\eta} = -a\frac{dx^i_\perp}{dt} = -\hat{p}^i_\perp, \tag{13.6}$$

where $\hat{\boldsymbol{p}}_\perp$ is the transverse part of the unit photon momentum vector and the last equality follows from Eq. (3.34) (and $p/E = 1$ for photons). This yields our desired integral for $\theta^i_S$:

$$\theta^i_S = \frac{x^i_\perp}{\chi} = -\frac{1}{\chi}\int_0^\chi \hat{p}^i_\perp(\chi'')d\chi''. \tag{13.7}$$

Note that we use $\chi''$ (and soon $\chi'$) as integration variables, not to be confused with conformal time derivatives. Now we can use the component of the geodesic equation derived in Eq. (3.72):

$$\frac{d\hat{p}^i}{dt} = \frac{1}{a}\left[\delta^{ik} - \hat{p}^i\hat{p}^k\right](\Phi - \Psi)_{,k}\,. \tag{13.8}$$

The factor $[\delta^{ik} - \hat{p}^i\hat{p}^k]$ is precisely the projection on directions transverse to $\boldsymbol{p}$, which in our small-angle approximation is transverse to the $z$-axis. Thus,

$$\frac{d\hat{p}^i_\perp}{d\chi} = -a\frac{d\hat{p}^i_\perp}{dt} = -a\left[\frac{1}{a}(\Phi - \Psi)_{,i}\right] = -2\Phi_{,i}. \tag{13.9}$$

The last equality holds since in the late universe there is no anisotropic stress and hence $\Phi = -\Psi$. Integrating Eq. (13.9) yields

$$\hat{p}^i_\perp(\chi'') = -2\int_0^{\chi''} d\chi'\Phi_{,i}(\boldsymbol{x}(\boldsymbol{\theta}, \chi'), \eta_0 - \chi') + C^i, \tag{13.10}$$

where

$$\boldsymbol{x}(\boldsymbol{\theta}, \chi') = \left(\chi'\theta^1,\ \chi'\theta^2,\ \chi'\right) \tag{13.11}$$

is the unperturbed photon path at which the potential is evaluated. That is, we evaluate the perturbation along the zeroth-order path, as the difference would be a second-order term in the deflection. We will fix the constant $C^i$ momentarily. Plugging in Eq. (13.10) into Eq. (13.7), we get

$$\theta^i_S = \frac{2}{\chi}\int_0^\chi d\chi''\int_0^{\chi''} d\chi'\Phi_{,i}(\boldsymbol{x}(\boldsymbol{\theta}, \chi'), \eta_0 - \chi') - C^i. \tag{13.12}$$

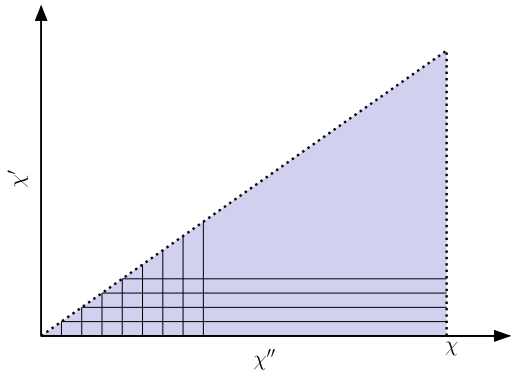



**FIGURE 13.5** Range of integration in the double integral of Eq. (13.13). The shaded region can be expressed as $0 < \chi' < \chi''$, $0 < \chi'' < \chi$ (performing the integral along the vertical lines) or as $\chi' < \chi'' < \chi$, $0 < \chi' < \chi$ (performing the integral along the horizontal lines). The latter is more convenient here, since the $\chi''$ integral is then trivial.

To determine $C^i$, we consider the limit of no deflections, $\Phi_{,i} \to 0$. Then, $\theta_S^i = -C^i$, which has to reduce to $\theta^i$, i.e. the true source position is equal to the observed position. Hence, $-C^i = \theta^i$ and we have

$$\theta_S^i = \theta^i + \frac{2}{\chi}\int_0^\chi d\chi'' \int_0^{\chi''} d\chi' \Phi_{,i}(\boldsymbol{x}(\boldsymbol{\theta}, \chi'), \eta_0 - \chi'). \tag{13.13}$$

The sign here is correct: An overdensity centered at $\boldsymbol{x}_\perp = 0$ has $\Phi > 0$ there, and therefore the derivative of $\Phi$ with respect to $x$ ($\Phi_{,i}$ with $i = 1$) is negative for $x > 0$. As such, the bending angle of a light ray passing the overdensity on the positive $x$-axis is negative, i.e., inward toward the overdensity, as we expect.

Now, the double integral in the $\chi'$, $\chi''$ plane is restricted to the shaded region in Fig. 13.5, so we can change orders of integration with the $\chi''$ integral ranging from $\chi'$ to $\chi$. The $\chi''$ integral is then trivial (since $\Phi_{,i}$ depends only on $\chi'$) and yields $\chi - \chi'$, so

$$\begin{aligned} \theta_S^i &= \theta^i + \Delta\theta^i \\ \Delta\theta^i(\boldsymbol{\theta}) &= 2\int_0^\chi d\chi' \Phi_{,i}(\boldsymbol{x}(\boldsymbol{\theta}, \chi'))\left(1 - \frac{\chi'}{\chi}\right). \end{aligned} \tag{13.14}$$

Using the fact that $\partial/\partial x^i = \chi'^{-1}\partial/\partial\theta^i$ under the integral via Eq. (13.11), we can write the deflection angle as the derivative of a lensing potential $\phi_{\rm L}$ on the sky,

$$\Delta\theta^i(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta^i}\phi_{\rm L}(\boldsymbol{\theta}) \tag{13.15}$$

$$\phi_{\rm L}(\boldsymbol{\theta}) \equiv 2\int_0^\chi \frac{d\chi'}{\chi'}\Phi(\boldsymbol{x}(\boldsymbol{\theta}, \chi'))\left(1 - \frac{\chi'}{\chi}\right). \tag{13.16}$$

The lensing potential is a weighted integral over $2\Phi$ along the photon path which, at the order we work in here, can be taken to be the unperturbed path Eq. (13.11). Notice that the

contribution to lensing by over- and underdensities close to the source are suppressed, since $1-\chi'/\chi$ goes to zero as $\chi'$ approaches $\chi$.

For galaxy shapes, we will also need the first derivative of the deflection angle, or equivalently the second-derivative matrix (*distortion tensor*) of the lensing potential:

$$\psi_{ij} \equiv \frac{\partial \Delta\theta^i}{\partial\theta^j} = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\phi_{\rm L}(\boldsymbol{\theta}) = 2\int_0^\chi d\chi' \Phi_{,ij}(\boldsymbol{x}(\boldsymbol{\theta},\chi'))\chi'\left(1-\frac{\chi'}{\chi}\right), \tag{13.17}$$

where we again converted a derivative with respect to $\boldsymbol{\theta}$ into a derivative with respect to $\boldsymbol{x}$ under the integral. The last equality will prove useful to relate lensing to the matter distribution.

## 13.3 CMB lensing

Gravitational lensing was first conceived as being applicable to light from discrete objects such as galaxies, and we will turn to this effect in the next section. Much later, cosmologists realized that a diffuse field such as the CMB will also be lensed. In fact, CMB lensing requires less formalism to describe, so we break with historical chronology here to walk through the impact of lensing on the CMB first.

For the CMB, the directional dependence of the intensity in Eq. (13.3) is contained in the temperature. Therefore, the equality translates into an equality between the observed temperature at position $\boldsymbol{\theta}$ and the unlensed temperature at $\boldsymbol{\theta}_S = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$. Taylor expanding leads to

$$\begin{aligned} T_{\rm obs}(\boldsymbol{\theta}) &= T_{\rm true}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}[\boldsymbol{\theta}]) \\ &\simeq T_{\rm true}(\boldsymbol{\theta}) + \Delta\theta^i \frac{\partial}{\partial\theta^i} T_{\rm true}(\boldsymbol{\theta}) + \frac{1}{2}\Delta\theta^i\,\Delta\theta^j \frac{\partial^2}{\partial\theta^i\partial\theta^j} T_{\rm true}(\boldsymbol{\theta}), \end{aligned} \tag{13.18}$$

where we have expanded up to second order in small deflections (we will soon see why), and the source distance is $\chi = \chi_*$, the distance to the last-scattering surface. We can now derive the statistics of $T_{\rm obs}$ from those of $T_{\rm true}$ and of $\phi_{\rm L}$ following a similar approach as for the nonlinear matter density in Sect. 12.2.

First, notice that the mean temperature of the CMB is unchanged by lensing. This is clear, since lensing only shuffles the arrival directions around without changing the surface brightness. So we can divide both sides of Eq. (13.18) by the mean temperature and subtract 1, and work with $\Theta_{\rm obs}$ and $\Theta_{\rm true} \equiv \Theta$ (in the following, we will drop the subscript "true" on the unlensed $\Theta$ for clarity; indeed, this is the $\Theta$ whose evolution we calculated in Ch. 9). Let us then switch to multipole space, in the flat-sky approximation throughout, and use

$$\Delta\boldsymbol{\theta}(\boldsymbol{l}) = i\boldsymbol{l}\phi_{\rm L}(\boldsymbol{l}), \tag{13.19}$$

the Fourier-space version of Eq. (13.15). Eq. (13.18) becomes

$$\Theta_{\rm obs}(\boldsymbol{l}) = \Theta(\boldsymbol{l}) - \int \frac{d^2l_1}{(2\pi)^2} \int \frac{d^2l_2}{(2\pi)^2} (2\pi)^2 \delta_{\rm D}^{(2)}(\boldsymbol{l}_1 + \boldsymbol{l}_2 - \boldsymbol{l})\, \boldsymbol{l}_1 \cdot \boldsymbol{l}_2 \phi_{\rm L}(\boldsymbol{l}_1)\, \Theta(\boldsymbol{l}_2)$$
$$+ \frac{1}{2} \int \frac{d^2l_1}{(2\pi)^2} \int \frac{d^2l_2}{(2\pi)^2} \int \frac{d^2l_3}{(2\pi)^2} (2\pi)^2 \delta_{\rm D}^{(2)}(\boldsymbol{l}_1 + \boldsymbol{l}_2 + \boldsymbol{l}_3 - \boldsymbol{l})$$
$$\times (\boldsymbol{l}_1 \cdot \boldsymbol{l}_3) \phi_{\rm L}(\boldsymbol{l}_1)\, (\boldsymbol{l}_2 \cdot \boldsymbol{l}_3) \phi_{\rm L}(\boldsymbol{l}_2)\, \Theta(\boldsymbol{l}_3). \tag{13.20}$$

We can now compute the angular power spectrum of $\Theta_{\rm obs}(\boldsymbol{l})$ in analogy to the derivation leading from the second- and third-order contributions to the matter density field in perturbation theory (Eq. (12.40)) to the next-to-leading order matter power spectrum Eq. (12.48). The result consists of coupling two quadratic terms and of coupling a linear term $\Theta(\boldsymbol{l})$ with a cubic-order one. The two differences to the case of the nonlinear matter density are that, first, we work in two rather than three dimensions; second, $\phi_{\rm L}$ and $\Theta$ are uncorrelated. This is because the contribution to $\phi_{\rm L}$ coming from near the source (where $\Theta$ originates) is suppressed, as we have seen. This fact actually simplifies the derivation.

In Exercise 13.2, you can then show that the power spectrum of $\Theta_{\rm obs}$ becomes

$$C^{\rm obs}(l) = C(l) + C^{(22)}(l) + 2C^{(13)}(l) \tag{13.21}$$
$$C^{(22)}(l) = \int \frac{d^2l_1}{(2\pi)^2} \left[\boldsymbol{l}_1 \cdot (\boldsymbol{l} - \boldsymbol{l}_1)\right]^2 C_{\phi_{\rm L}\phi_{\rm L}}(l_1) C(|\boldsymbol{l} - \boldsymbol{l}_1|).$$
$$C^{(13)}(l) = -\frac{1}{4} \left[ \int \frac{d^2l_1}{(2\pi)^2} l_1^2 C_{\phi_{\rm L}\phi_{\rm L}}(l_1) \right] l^2 C(l)$$

Here, the power spectrum of the lensing potential is defined through

$$\left\langle \phi_{\rm L}(\boldsymbol{l}) \phi_{\rm L}^*(\boldsymbol{l}') \right\rangle = (2\pi)^2 \delta_{\rm D}^{(2)}(\boldsymbol{l} - \boldsymbol{l}') C_{\phi_{\rm L}\phi_{\rm L}}(l). \tag{13.22}$$

We defer its derivation to Sect. 13.5 below, but it is shown in Fig. 13.7 (with a prefactor of $l^4/4$).

The (13) term in Eq. (13.21) corresponds to a damping of power: we can write $C(l) + 2C^{(13)}(l) = (1 - l^2/l_{\rm lens}^2) C(l)$. The factor $l_{\rm lens}^{-2}$ is proportional to the RMS displacement squared, so we can interpret this as smoothing of the observed temperature distribution on the sky due to random lensing deflections. The (22) term on the other hand convolves the unlensed power spectrum with the power spectrum of lensing deflections. This effectively leads to a damping of the peaks in the $C(l)$. It also adds power to the CMB anisotropies on very small scales: we have seen in Ch. 9 that when $l \gg l_D$, the angular damping scale, the primary CMB anisotropies are exponentially suppressed. The integral in the (22) term in Eq. (13.21), however, transfers power from large to small scales, so that the lensed temperature anisotropies are actually larger than the primordial ones in the damped regime.

Both effects are visible in Fig. 13.6, which shows both lensed and primordial temperature power spectra (CMB lensing is incorporated in Boltzmann codes such as CAMB and CLASS). But they are really evident only when looking at the fractional difference (bottom

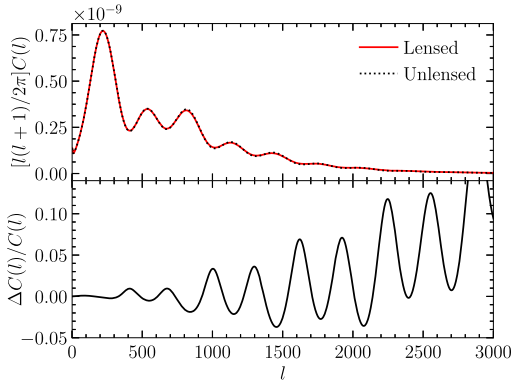



**FIGURE 13.6** Effect of lensing on the CMB temperature power spectrum. The top panel shows the power spectrum before ($C(l)$, dashed) and after lensing ($C^{\rm obs}(l)$, Eq. (13.21), solid). The bottom panel shows the fractional difference between these very similar spectra. Lensing smoothes out the peaks in the power spectrum, and adds power on very small scales.

panel). The smoothing of the peaks leads to an oscillatory pattern with opposite sign than that in the $C(l)$ itself. The increase of small-scale power in $C^{\rm obs}(l)$ is also visible. Despite its small size, the lensing effect is larger than the error bars in current measurements of the $C(l)$ and needs to be accounted for.

The effect on the power spectrum depicted in Fig. 13.6 is important but it is actually not the most exciting aspect of CMB lensing. In particular, it does not allow us to easily extract the desired power spectrum of the lensing potential $\phi_{\rm L}$. There is, however, a much more powerful way of invoking Eq. (13.18) to accomplish this goal. In Eq. (13.21) we only considered the case $\langle\Theta_{\rm obs}(\boldsymbol{l})\Theta^*_{\rm obs}(\boldsymbol{l}')\rangle$ with $\boldsymbol{l}'=\boldsymbol{l}$. In case of the unlensed CMB, we argued in Ch. 9 that these are the only correlations that are nonzero. This is no longer true for the lensed CMB: Eq. (13.18) leads to nonzero correlations between the off-diagonal modes, i.e. those with $\boldsymbol{l}'\neq\boldsymbol{l}$. So let us now consider the correlation of two $\Theta_{\rm obs}$ with $\boldsymbol{l}'\neq\boldsymbol{l}$. Using the first line of Eq. (13.20), this yields

$$\begin{aligned}\langle\Theta_{\rm obs}(\boldsymbol{l})\Theta^*_{\rm obs}(\boldsymbol{l}')\rangle\Big|_{\phi_{\rm L}} &\overset{\boldsymbol{l}'\neq\boldsymbol{l}}{=} -\int\frac{d^2l_1}{(2\pi)^2}\Big[\phi^*_{\rm L}(\boldsymbol{l}_1)\boldsymbol{l}_1\cdot(\boldsymbol{l}'-\boldsymbol{l}_1)\langle\Theta(\boldsymbol{l})\Theta^*(\boldsymbol{l}'-\boldsymbol{l}_1)\rangle\\ &\qquad+\phi_{\rm L}(\boldsymbol{l}_1)\boldsymbol{l}_1\cdot(\boldsymbol{l}-\boldsymbol{l}_1)\langle\Theta(\boldsymbol{l}-\boldsymbol{l}_1)\Theta^*(\boldsymbol{l}')\rangle\Big]\\ &=\phi_{\rm L}(\boldsymbol{l}-\boldsymbol{l}')\,(\boldsymbol{l}-\boldsymbol{l}')\cdot\big[\boldsymbol{l}C(l)-\boldsymbol{l}'C(l')\big],\end{aligned}\tag{13.23}$$

where the subscript $\phi_{\rm L}$ on the expectation value on the left-hand side signifies that we are taking the expectation value over the primary CMB fluctuations while keeping the lensing field $\phi_{\rm L}(\boldsymbol{l})$ fixed. In the second line we have used the fact that the correlations of the *primary* anisotropies $\Theta(\boldsymbol{l})$ are nonzero only when the $\boldsymbol{l}$ vectors in the argument are equal, as well as the reality condition of the lensing potential: $\phi^*_{\rm L}(\boldsymbol{l}_1)=\phi_{\rm L}(-\boldsymbol{l}_1)$.

Eq. (13.23) shows that, by combining different modes $\Theta_{\rm obs}(\boldsymbol{l})$, $\Theta^*_{\rm obs}(\boldsymbol{l}')$ at fixed $\boldsymbol{L} = \boldsymbol{l} - \boldsymbol{l}'$, we can reconstruct the lensing potential $\phi_{\rm L}(\boldsymbol{L})$. The idea is to take the weighting of modes that maximizes the signal-to-noise (Hu, 2001; Hu and Okamoto, 2002). Because the measurement of $\phi_{\rm L}$ involves two powers of $\Theta_{\rm obs}$, this technique is called *optimal quadratic estimator*. The result is a map of the projected matter density on the sky that can be correlated with other maps and whose statistics we can measure.

To gain some intuition for this estimator, consider the case when $|\boldsymbol{L}| = |\boldsymbol{l} - \boldsymbol{l}'| \ll l, l'$, so that the left-hand side of Eq. (13.23) is very close to a standard power spectrum configuration. Then, Eq. (13.23) describes the modulation of the observed small-scale CMB power spectrum by a long-wavelength lensing potential with wavenumber $\boldsymbol{L}$. This small-scale power spectrum is anisotropic due to the lensing effect (notice the factors of $\boldsymbol{L} \cdot \boldsymbol{l}$ on the right-hand side, which shows that the observed power spectrum depends on the angle of the small-scale mode with the $\boldsymbol{L}$-mode). The quadratic estimator uses this characteristic anisotropy to reconstruct the modulating lensing mode $\phi_{\rm L}(\boldsymbol{L})$.

Finally, we considered only the lensing of CMB temperature perturbations here. CMB lensing becomes even more powerful when combined with polarization. If the primary CMB has only $E$-mode polarization, then the reshuffling of positions due to lensing will generate $B$-modes (if you look at Fig. 10.3 carefully, you can tell that a reshuffling of positions will turn a pure $E$-mode into a mixture of $E$- and $B$-modes). Then, the small-scale lensing-generated $B$-modes can be used as an almost cosmic-variance-free channel for reconstructing $\phi_{\rm L}(\boldsymbol{L})$.

## 13.4 Galaxy shapes

Let us now switch from the CMB to galaxies. Eq. (13.3) shows that entire galaxy images are displaced from their true position to the observed one, similar to how the CMB temperature is remapped. Unfortunately, we do not know the true galaxy positions a priori. One possible approach then is to use the isotropy of the unlensed projected galaxy density field, which is closely analogous to the CMB lensing approach of Sect. 13.3. However, this approach is usually not nearly as powerful as the one we will describe here. Instead of looking at the overall displacement of the galaxy position, we use the relative displacement of different parts of the galaxy image, that is, the distortion in the observed shapes of the galaxies (as shown in extreme form for the arc-like images in Fig. 13.2). This happens because the lensing deflection angle varies on the sky, so it also varies across the image of a galaxy. In the simplest case, lensing turns the image of a circular galaxy into an elliptical one.

To describe this effect, then, we need to come up with quantitative measures of galaxy shapes, and see how these are affected by gravitational lensing. The simplest measure of a galaxy shape is the set of second moments (or quadrupole moments) of its image. Imagine centering an image at the origin $(\theta_x, \theta_y) = (0, 0)$. Then the second moments are defined as

$$q_{ij} \equiv \langle \theta_i \theta_j \rangle_{I_{\rm obs}} \equiv \frac{1}{F} \int d^2\theta \, I_{\rm obs}(\theta) \theta_i \theta_j \tag{13.24}$$

where angular brackets denote intensity-weighted averages over the image, and the second moments are normalized by

$$F = \int d^2\theta\, I_{\rm obs}(\boldsymbol{\theta}), \tag{13.25}$$

the total flux of the image. In particular, here we have chosen the origin such that $\langle\theta_x\rangle_{I_{\rm obs}} = \langle\theta_y\rangle_{I_{\rm obs}} = 0$. $q_{ij}$ is a symmetric $2\times 2$ matrix which we can write as

$$q_{ij} = \frac{1}{2}q\begin{pmatrix} 1+\epsilon_1 & \epsilon_2 \\ \epsilon_2 & 1-\epsilon_1 \end{pmatrix}. \tag{13.26}$$

The three independent components now are the trace $q = {\rm Tr}[q_{ij}]$ and $\epsilon_1$, $\epsilon_2$. A circular image has $\epsilon_1 = \epsilon_2 = 0$, while $\sqrt{q}$ provides a measure of the angular size of the image. Eq. (13.26) is very similar to the polarization tensor in Eq. (10.2): $q/2$ is equivalent to the intensity $I$, while $\epsilon_i$ are equivalent to the scaled $Q$ and $U$ polarizations:

$$\epsilon_1 \leftrightarrow \frac{Q}{I}; \quad \epsilon_2 \leftrightarrow \frac{U}{I}. \tag{13.27}$$

So we can interpret the polarization pattern shown in Fig. 10.2 equivalently as showing galaxy images described by different values of $\epsilon_1$ and $\epsilon_2$. It further means that we can immediately adopt the results of Sect. 10.1, and define the $E$- and $B$-modes of the galaxy ellipticity field. We will begin by assuming that galaxy shapes are intrinsically random, so that all observed shape correlations are due to lensing. This is not entirely correct, as we will see below.

So let us derive how lensing affects the galaxy shape tensor $q_{ij}$. A distortion due to lensing happens because the deflection angle $\Delta\theta$ varies across the galaxy image. Hence, we need the derivative of the observed position $\theta$ with respect to the observed angle. It is conventional to define the $2\times 2$ transformation matrix,

$$A_{ij} \equiv \frac{\partial\theta_S^i}{\partial\theta^j} = \begin{pmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{pmatrix}. \tag{13.28}$$

In the second equality we have assumed that $A_{ij}$ is symmetric, which allows us to write it in a form analogous to Eq. (13.26). This follows from the fact that the deflection angle can be written as the gradient of a scalar lensing potential (Eq. (13.16)). The antisymmetric part of $A_{ij}$, which thus vanishes at leading order, corresponds to an image rotation (see Exercise 13.5).

The quantity $\kappa$, the analogue of image size $q$ or intensity $I$, is called the *convergence*; it describes how the flux and size of an image are modified. The components we are most interested in are the two components of the *shear*,

$$\begin{aligned} \gamma_1 &= -\frac{A_{11}-A_{22}}{2}, \\ \gamma_2 &= -A_{12}. \end{aligned} \tag{13.29}$$

Again, the $\gamma_i$ are analogous to the $\epsilon_i$ and to the polarization $Q$ and $U$.

We can directly express $A_{ij}$ in terms of the distortion tensor defined in Eq. (13.17):

$$A_{ij} = \delta_{ij} + \psi_{ij}, \qquad \text{and} \qquad \psi_{ij} = \begin{pmatrix} -\kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & -\kappa + \gamma_1 \end{pmatrix}. \tag{13.30}$$

In other words, $\kappa$, $\gamma_1$ and $\gamma_2$ are well-defined functions of the integrated gravitational potential. We can now derive how they influence the shapes of galaxy images.

In the presence of lensing, the observed second moments of a given galaxy are, through Eq. (13.3) and Eq. (13.24), given by

$$q_{ij} = \frac{\int d^2\theta\, I_{\rm true}(\boldsymbol{\theta}_S)\theta_i\theta_j}{\int d^2\theta\, I_{\rm true}(\boldsymbol{\theta}_S)}. \tag{13.31}$$

The integrals here are over the observed angles $\boldsymbol{\theta}$, while the intensity depends on the angle from which the photon started at the source, $\boldsymbol{\theta}_S$. We now want to make use of the fact that the deflection angle $\Delta\boldsymbol{\theta}$ only varies slightly across the galaxy image. We expand

$$\theta^i_S(\boldsymbol{\theta}) = \theta^i + \Delta\theta^i + \frac{\partial\Delta\theta^i}{\partial\theta^j}\theta^j + \cdots = A^{ij}\theta^j + \Delta\theta^i + \cdots, \tag{13.32}$$

where $\Delta\theta^i$ and its derivative are to be evaluated at the galaxy centroid position. Keep in mind that $A_{ij}$ and $\Delta\theta^i$ are always evaluated at a fixed position, so we can pull them out of the integrals over the galaxy image. This approximation is sufficiently accurate for weak lensing, as higher-order corrections in $\theta$ become significant only on very small scales.

Let us drop the constant shift $\Delta\theta^i$ for the time being; this is just the shift in the galaxy's apparent position, so we do not expect it to affect the shape. Below we will show that this is indeed the case. Then, we can replace $\theta^i$ with $(A^{-1})^{ij}\theta_{S,j}$ everywhere in Eq. (13.31). To do the integrals, first change the integration variable to $\boldsymbol{\theta}_S$. This leads to

$$\begin{aligned} q_{ij} &= \frac{1}{F}\int d^2\theta_S \left|\frac{\partial\theta_k}{\partial\theta_{S,l}}\right| I_{\rm true}(\boldsymbol{\theta}_S)(A^{-1}\theta_S)_i(A^{-1}\theta_S)_j, \\ F &= \int d^2\theta_S \left|\frac{\partial\theta_k}{\partial\theta_{S,l}}\right| I_{\rm true}(\boldsymbol{\theta}_S). \end{aligned} \tag{13.33}$$

Let us begin with the second line, the total flux of the lensed image. The Jacobian factor is equal to $|A^{-1}| = |A|^{-1}$, and can be pulled out of the integral to obtain

$$F = |A|^{-1} F_{\rm true}, \tag{13.34}$$

where $F_{\rm true}$ is the flux that would be observed if the galaxy was not lensed, and the inverse determinant is given by (Exercise 13.3)

$$\mu \equiv |A|^{-1} = \frac{1}{(1-\kappa)^2 - \gamma_1^2 - \gamma_2^2}. \tag{13.35}$$

This quantity, which describes the increase or decrease in total galaxy flux, is called *magnification*. Since lensing conserves $I_\nu$, i.e. surface brightness, the increase in flux of an image is equal to the increase in area. The magnification has many important ramifications in all areas of lensing, and can be measured through its effect on galaxy number counts (see Exercise 13.11) as well as galaxy size correlations. Eq. (13.35) is in fact valid beyond the weak-lensing regime.

For the shape tensor in the first line of Eq. (13.33), the Jacobian factors cancel and it becomes

$$q_{ij} = (A^{-1})_i{}^k (A^{-1})_j{}^l q_{kl}^{\rm true}, \tag{13.36}$$

where $q_{kl}^{\rm true}$ is the second-moment tensor that would be observed in the absence of lensing. This is the transformation law for the second moments under lensing.

Now, if we include the constant $\Delta\boldsymbol{\theta}$ that we have dropped after Eq. (13.32), we obtain linear and quadratic terms in $\Delta\boldsymbol{\theta}$. The latter is higher order in our treatment, so we can immediately drop it. The linear-order terms on the other hand are proportional to $\int d^2\theta_S I_{\rm true}(\boldsymbol{\theta}_S)\theta_S^i$, which vanishes since we have chosen the coordinates such that the origin is the centroid of the image, and so the dipole of the intensity vanishes. Thus, we were justified in dropping $\Delta\theta^i$.

We now proceed by linearizing Eq. (13.36) in $\kappa$ and $\gamma_i$, and using Eq. (13.30). This yields

$$q_{ij} \overset{\rm linear}{=} q_{ij}^{\rm true} - \psi_i{}^k q_{kj}^{\rm true} - \psi_j{}^l q_{il}^{\rm true}. \tag{13.37}$$

Using Eq. (13.26) for both $q_{ij}$ and $q_{ij}^{\rm true}$, it is now straightforward to derive the transformation of the trace $q$ and the ellipticity components $\epsilon_i$ under lensing. We will only give the result here, leaving the intermediate steps as Exercise 13.3. First, the trace of $q_{ij}$ is

$$q = {\rm Tr}\, q_{ij} = q_{\rm true}\left[1 + 2\kappa + 2(\epsilon_1^{\rm true}\gamma_1 + \epsilon_2^{\rm true}\gamma_2)\right]. \tag{13.38}$$

Since $q$ is a measure for the area of the galaxy image, which is proportional to the flux when surface brightness is kept fixed, we expect it to increase as $\mu \simeq 1 + 2\kappa$. Neglecting the higher-order ellipticity corrections, this is precisely what Eq. (13.38) dictates. For the ellipticities, we have

$$\begin{aligned}
\epsilon_1 &= \frac{q_{11} - q_{22}}{q} = \left(1 - 2[\epsilon_1^{\rm true}\gamma_1 + \epsilon_2^{\rm true}\gamma_2]\right)\epsilon_1^{\rm true} + 2\gamma_1, \\
\epsilon_2 &= \frac{2q_{12}}{q} = \left(1 - 2[\epsilon_1^{\rm true}\gamma_1 + \epsilon_2^{\rm true}\gamma_2]\right)\epsilon_2^{\rm true} + 2\gamma_2.
\end{aligned} \tag{13.39}$$

Notice that here we have assumed that $\kappa$, $\gamma_1$, $\gamma_2$ are small, but not that $\epsilon_i$ or $\epsilon_i^{\rm true}$ are small, since in reality galaxies are not close to circular in general. Nevertheless, if we also assume that the $\epsilon_i^{\rm true}$ are small, then the result becomes very simple:

$$\epsilon_i = \epsilon_i^{\rm true} + 2\gamma_i \qquad (\epsilon_i^{\rm true} \ll 1). \tag{13.40}$$

By measuring ellipticities of distant galaxies, therefore, we can get an estimate of the shear field, which depends directly on the gravitational potential via Eq. (13.17). In fact, we do not know the true ellipticity components of individual galaxies. We do know their distribution, however,[1] which is fairly narrow with an RMS width of typically $\langle(\epsilon_1^{\rm true})^2 + (\epsilon_2^{\rm true})^2\rangle^{1/2}/\sqrt{2} \simeq 0.3$. By summing the ellipticities of many galaxies in a small area (pixel) of sky, the noise due to the random intrinsic ellipticities averages out and the lensing signal remains. The shear fields $\gamma$ whose statistics we derive next should be understood as being measured in this way.

## 13.5 Weak-lensing statistics

On average, each of the components of the distortion tensor is zero: $\langle\psi_{ij}\rangle = 0$, since $\langle\phi_{\rm L}\rangle = 0$. To make our money, therefore, we need to do just what we did for the CMB and galaxy distributions: compute either the angular correlation function or its Fourier transform, the power spectrum. We will do this in the flat-sky approximation again, well justified since most of the lensing signal is on small scales (angular wavenumbers $l \gtrsim 100$). Then, the derivation essentially amounts to combining the results of Sect. 11.2 (projected angular correlations) and Sect. 10.1 (since shear is analogous to polarization). Consequently, we begin with the power spectrum in Fourier space, before deriving the real-space correlation functions. We will then turn to other important lensing statistics, in particular the cross-correlation with galaxy counts.

### 13.5.1 Shear power spectrum

In order to derive the shear power spectrum, we take the two-dimensional Fourier transform of Eq. (13.17):

$$-\psi_{ij}(\boldsymbol{l}) = l_i l_j \phi_{\rm L}(\boldsymbol{l}). \tag{13.41}$$

This relation allows us to directly compute the power spectra of shear and convergence, by making use of Eq. (13.30). Given this expression for the distortion tensor, which we can treat just like $I_{ij}$ in Sect. 10.1, the $E$-mode is given via Eq. (10.6) (notice that we have to subtract the trace from $-\psi_{ij}$)

$$E(\boldsymbol{l}) = \left(\frac{l^i l^j}{l^2} - \frac{1}{2}\delta_{ij}\right)[-\psi_{ij}(\boldsymbol{l})] = \frac{1}{2}l^2\phi_{\rm L}(\boldsymbol{l}) = \kappa(\boldsymbol{l}). \tag{13.42}$$

The last equality follows from Eq. (13.30) as well, since taking the trace of $-\psi_{ij}$ yields $2\kappa = -(\partial^2/\partial\theta_1^2 + \partial^2/\partial\theta_2^2)\phi_{\rm L}$. The $B$-mode vanishes. This can be derived through Eq. (10.5). Alternatively, one can show that the $B$-mode of the distortion tensor can, at linear order, only be sourced by a curl-type deflection angle, whereas Eq. (13.14) clearly is of gradient type (Exercise 13.5).

[1] Since lensing is a small effect, to first order we can use the distribution of observed shapes of all galaxies as a proxy for the distribution of intrinsic shapes.

Deriving the power spectrum of the $E$-mode and therefore of $\kappa$ now follows:

$$\left\langle E(\boldsymbol{l})E^*(\boldsymbol{l}')\right\rangle = (2\pi)^2\delta_{\rm D}^{(2)}(\boldsymbol{l}-\boldsymbol{l}')C_{EE}(l), \tag{13.43}$$

where the power spectra are related by

$$C_{EE}(l) = C_{\kappa\kappa}(l) = \frac{1}{4}l^4 C_{\phi_{\rm L}\phi_{\rm L}}(l). \tag{13.44}$$

Conveniently, $\phi_{\rm L}$ is a scalar on the sky, so its power spectrum can be computed exactly as we did for the angular galaxy correlations in Sect. 11.2. So let us write, in analogy with Eq. (11.39),[2]

$$\phi_{\rm L}(\boldsymbol{\theta}) = 2\int_0^\infty \frac{d\chi}{\chi} g_{\rm L}(\chi)\,\Phi\left(\boldsymbol{x}(\chi), \eta_0-\chi\right). \tag{13.45}$$

The only difference to Eq. (11.39) is the different kernel $g_{\rm L}(\chi)$, which we will specify momentarily, and that we have to replace $\delta_{\rm g}$ with $2\Phi$. The derivation then proceeds just as in Sect. 11.2, and we obtain from Eq. (11.47)

$$C_{\phi_{\rm L}\phi_{\rm L}}(l) = 4\int_0^\infty \frac{d\chi}{\chi^2}\frac{g_{\rm L}^2(\chi)}{\chi^2} P_\Phi\left(k=\frac{l+1/2}{\chi}, \eta(\chi)\right). \tag{13.46}$$

Now let us derive the projection kernel $g_{\rm L}$ for $\phi_{\rm L}$, for which we need to include one more observational complication. Eq. (13.17) gives the distortion tensor for a galaxy at fixed distance $\chi$. Because it is a small effect and requires large statistics, weak lensing is usually measured in photometric surveys which do not yield a distance for each galaxy. So instead, we are measuring the statistics of the distortion tensor for a distribution of galaxy redshifts. Let us call this distribution $W(\chi)$, just as we did when studying angular correlations in Sect. 11.2. Again, let us normalize $W$ so that $\int_0^\infty d\chi\, W(\chi) = 1$. Then, the lensing potential $\phi_{\rm L}$ (Eq. (13.16)) is

$$\phi_{\rm L}(\boldsymbol{\theta}) = 2\int_0^\infty d\chi\, W(\chi)\int_0^\chi \frac{d\chi'}{\chi'}\Phi(\boldsymbol{x}(\boldsymbol{\theta},\chi'))\left(1-\frac{\chi'}{\chi}\right). \tag{13.47}$$

We can again simplify this double integral by changing orders of integration (almost exactly as depicted in Fig. 13.5). Then

$$\phi_{\rm L}(\boldsymbol{\theta}) = 2\int_0^\infty \frac{d\chi'}{\chi'} g_{\rm L}(\chi')\Phi(\boldsymbol{x}(\boldsymbol{\theta},\chi')) \tag{13.48}$$

where

$$g_{\rm L}(\chi') \equiv \int_{\chi'}^\infty d\chi\left(1-\frac{\chi'}{\chi}\right)W(\chi). \tag{13.49}$$

[2] We have replaced $\hat{\boldsymbol{n}}$ with $\boldsymbol{\theta}$ here, since we are working in the flat-sky approximation from the outset.

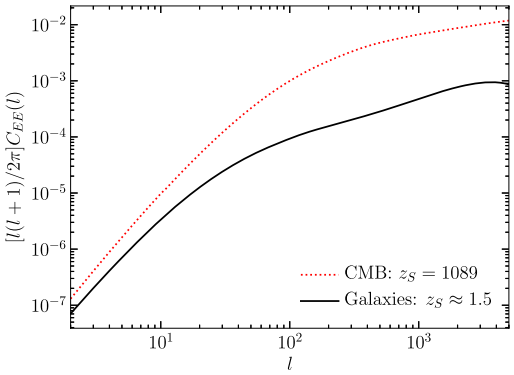



**FIGURE 13.7** The $E$-mode shear power spectrum for the CMB (red, dashed, $z_S = 1089$), and for source galaxies with a Gaussian redshift distribution centered at $z_S = 2$ with an RMS width of $\Delta z = 0.2$ (black, solid). Note that $C_{EE}(l) = C_{\kappa\kappa}(l) = (l^4/4)C_{\phi_L\phi_L}(l)$ in our flat-sky treatment. Both curves are computed using a prescription for the nonlinear matter power spectrum as a function of redshift calibrated on simulations (see Sect. 12.7). The power spectrum is larger for CMB lensing than for galaxies, because of the longer distance lever arm.

We are essentially done, but let us massage the power spectra into slightly simpler forms. First, assume that $l$ is sufficiently large so we can neglect the $+1/2$ in the argument of $P_\Phi$. Second, we can use the Poisson equation (8.6) to relate $\Phi$ to the nonlinear matter density perturbation $\delta_m$ (which, as we argued in Ch. 12, remains valid for nonlinear structure):

$$k^2\Phi = -\frac{3}{2}\Omega_m H_0^2 a^{-1}\delta_m, \qquad \text{so that} \qquad P_\Phi\left(k = \frac{l}{\chi}\right) = \left(\frac{3\Omega_m H_0^2}{2a}\right)^2 \frac{\chi^4}{l^4} P\left(\frac{l}{\chi}\right). \tag{13.50}$$

Then, the factor of $\chi^4$ cancels the denominators in Eq. (13.46) and the factor of $l^4$ cancels in Eq. (13.44), so we are left with

$$C_{EE}(l) = C_{\kappa\kappa}(l) = \left(\frac{3}{2}\Omega_m H_0^2\right)^2 \int_0^\infty d\chi\, a^{-2}(\chi) g_L^2(\chi) P\left(k = \frac{l}{\chi}, \eta(\chi)\right). \tag{13.51}$$

That is, the angular power spectrum of galaxy ellipticity correlations induced by lensing is directly proportional to an integral over the nonlinear matter power spectrum $P(k)$, weighted by the lensing kernel Eq. (13.49). Notice that $\Omega_m$ enters, since the amplitude of gravitational lensing is controlled by the potential perturbations, which depend on the total amount of matter sourcing gravity and hence the mean matter density. The result in our fiducial $\Lambda$CDM cosmology is shown in Fig. 13.7. Clearly, most of the shear signal is on small angular scales.

Our results now allow us to understand the power of weak lensing. First, it probes the nonlinear matter power spectrum $P(k)$ directly, without the annoying bias factors in between that we have for galaxies. It can then be used to constrain the amplitude $\sigma_8$ of fluctuations as well as $\Omega_m$. It is important to note that Eq. (13.51) remains quite accurate

even on small scales, where $P(k)$ is probed in the nonlinear regime. That is, the nonlinearities in the shear, which we have neglected throughout, are much smaller than those in the three-dimensional density field. Our theory predictions for weak lensing then are limited only by our ability to predict the nonlinear matter power spectrum, which becomes sensitive to the effects of baryons on very small scales (Sect. 12.3). Second, weak lensing also tells us about the expansion history, since the amplitude of $C_{EE}(l)$ depends on the comoving distance $\chi$ to the source galaxies, i.e. it depends on the distance-redshift relation. Further, we know that there are two components to galaxy ellipticities, but only one of them, the $E$-mode is nonzero. Thus, the observed $B$-mode can be used as a rigorous test for systematics in the measurement.

We should also mention some of the challenges of lensing. First, the signal is small (see the vertical axis in Fig. 13.7 and Exercise 13.8), because photons travel at the speed of light and are hard to deflect. This means we need a large number of galaxies to measure it precisely, which in turn means that most of these galaxies will be faint. This poses challenges for robust measurements of shapes. Second, in practice we do not know the redshift distributions of the source galaxies accurately, so they must be calibrated using external measurements or the lensing statistics themselves (at the price of losing some of the cosmological constraining power).

Finally, galaxy shapes are not entirely uncorrelated intrinsically, and a model for their intrinsic correlations, known as *intrinsic alignments*, has to be included in the interpretation of the measurement. On large scales, the leading effect that intrinsically correlates galaxy shapes is the tidal field. So, writing a linear relation between the shape and the second-derivative tensor of the gravitational potential, we have

$$q_{ij}^{\mathrm{IA}}(\boldsymbol{x},\eta) = c_1(\eta)\frac{\partial^2}{\partial\theta^i\partial\theta^j}\Psi(\boldsymbol{x},\eta) = -c_1(\eta)\chi^2\Phi_{,ij}(\boldsymbol{x},\eta), \tag{13.52}$$

where $c_1$ only depends on time. This is the analogue for shapes of the linear bias relation for the galaxy density. Since the shape is described by a $2\times 2$ symmetric matrix on the sky, we have to write the tidal field in a form consistent with $q_{ij}$. In Exercise 13.9 you can work out how this effect contributes to shear statistics. Observationally, tidal alignment has been detected to high significance, but the coefficient $c_1$ is generally less than one. That is, galaxies are less aligned than they are clustered. Nevertheless, given the high precision of current weak-lensing measurements, this effect needs to be incorporated. Note that we do not expect Eq. (13.52) to be sufficient on mildly or fully nonlinear scales; rather, nonlinear alignment terms will become important in analogy to the nonlinear bias contributions derived in Sect. 12.6.

### 13.5.2 Shear correlation function

We now turn to the Fourier transform of the shear power spectrum, the correlation function. In the large-scale structure realm, unlike the CMB, correlation functions are frequently employed because they are simpler to measure than power spectra. The

$E/B$-decomposition we have used so far naturally works in Fourier space. So there is no simple notion of an "$E$-mode correlation function." Nevertheless, the absence of $B$-modes also manifests itself in the shear correlation functions.

To begin, we again use the polarization analogy to derive, from Eq. (10.11),

$$\begin{pmatrix} \gamma_1(\boldsymbol{l}) \\ \gamma_2(\boldsymbol{l}) \end{pmatrix} = \begin{pmatrix} \cos 2\phi_l \\ \sin 2\phi_l \end{pmatrix} E(\boldsymbol{l}), \tag{13.53}$$

where we have dropped any $B$-modes from the outset (see Exercise 13.10). Let us compute then the auto-correlation of $\gamma_1$:

$$\begin{aligned} \langle \gamma_1(\mathbf{0})\gamma_1(\boldsymbol{\theta})\rangle &= \int \frac{d^2l}{(2\pi)^2} \int \frac{d^2l'}{(2\pi)^2} \cos 2\phi_l \cos 2\phi_{l'} \left\langle E(\boldsymbol{l})E(\boldsymbol{l}')\right\rangle e^{i\boldsymbol{l}\cdot\boldsymbol{\theta}} \\ &= \int \frac{d^2l}{(2\pi)^2} e^{il\theta\cos\phi_l} \cos^2 2\phi_l \, C_{EE}(l), \end{aligned} \tag{13.54}$$

where we have aligned the $l_x$-axis with $\boldsymbol{\theta}$. We will return to this below. The correlation function of $\gamma_2$ follows the same expression with $\cos^2 2\phi_l$ replaced by $\sin^2 2\phi_l$. So let us take the sum of the two, since that will be simple to compute:

$$\langle \gamma_1(\mathbf{0})\gamma_1(\boldsymbol{\theta})\rangle + \langle \gamma_2(\mathbf{0})\gamma_2(\boldsymbol{\theta})\rangle = \int \frac{ldl}{2\pi} J_0(l\theta) C_{EE}(l), \tag{13.55}$$

where we have used Eq. (C.24). On the other hand, the difference between the two is also straightforward to compute

$$\begin{aligned} \langle \gamma_1(\mathbf{0})\gamma_1(\boldsymbol{\theta})\rangle - \langle \gamma_2(\mathbf{0})\gamma_2(\boldsymbol{\theta})\rangle &= \int \frac{d^2l}{(2\pi)^2} e^{il\theta\cos\phi_l} \cos^2 4\phi_l \, C_{EE}(l) \\ &= \int \frac{ldl}{2\pi} J_4(l\theta) C_{EE}(l), \end{aligned} \tag{13.56}$$

again via Eq. (C.24).

Now, the components $\gamma_{1,2}$ depend on the definition of the coordinate system. We have aligned $\boldsymbol{\theta}$ with the $x$-axis, so the separation between the two galaxies whose shape we are correlating is horizontal in the illustration of Fig. 10.2. Therefore, nonzero $\gamma_1$ corresponds to galaxies either aligned with $\boldsymbol{\theta}$ or orthogonal to it (just like $Q$-polarization). Similarly, nonzero $\gamma_2$ corresponds to galaxies oriented at 45° from the separation vector $\boldsymbol{\theta}$ (like $U$-polarization). We can generalize this decomposition so that the results are independent of the coordinate definition. Functionally, doing that requires not a measurement of $\gamma_1$ or $\gamma_2$, but rather $\gamma_t$ and $\gamma_\times$: the *tangential shear* $\gamma_t$ is the component of the shear that is parallel or perpendicular to the line connecting two given galaxies (it is negative for a shear parallel to this line). The *cross-component* of the shear is oriented at 45° or 135° with respect to the separation vector. The above results then apply to these coordinate-invariant definitions:

$$\langle \gamma_t(\mathbf{0})\gamma_t(\boldsymbol{\theta})\rangle \pm \langle \gamma_\times(\mathbf{0})\gamma_\times(\boldsymbol{\theta})\rangle = \xi_{+,-}(\theta) \tag{13.57}$$

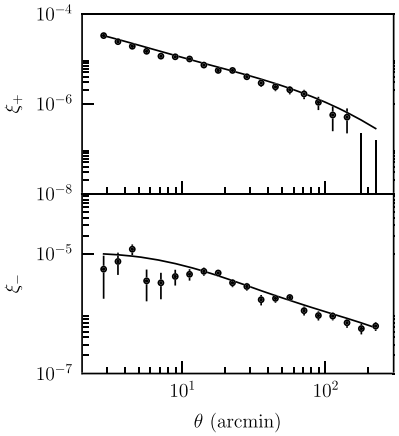



**FIGURE 13.8** Measurement of the shear correlation functions from the first year of data from the Dark Energy Survey. This includes the complete source galaxy sample, not split up by photometric redshift. The solid line shows the prediction for the best-fitting $\Lambda$CDM model (including intrinsic alignments and other effects). From Troxel et al. (2018).

where

$$\xi_{+,-}(\theta) = \int \frac{ldl}{2\pi} J_{0,4}(l\theta) C_{EE}(l) \tag{13.58}$$

are the "plus-" and "minus-" type shear correlation functions, which are both uniquely determined by $C_{EE}(l) = C_{\kappa\kappa}(l)$. You can also show that the cross-correlation function between the two shear components vanishes: $\langle \gamma_t(\mathbf{0})\gamma_\times(\boldsymbol{\theta})\rangle = 0$ (this is quite simple when aligning $\boldsymbol{\theta}$ with the $x$-axis again). All of these relations really follow from the fact that the ellipticity field is composed purely of $E$-modes; if present, $B$-modes also contribute to $\xi_+$ and $\xi_-$, but with a minus sign in the latter, as you can show in Exercise 13.10.

Fig. 13.8 shows a measurement from the Dark Energy Survey. Clearly, lensing is detected at very high signal-to-noise, especially on small scales (toward the left in the figure). The measurement is entirely compatible with the $\Lambda$CDM prediction. The actual cosmology constraints were obtained from a larger data vector: the source galaxy sample was split into four subsamples based on their photometric redshift estimates. Then, all shape auto- and cross-correlations between these different samples were used, yielding more information on the expansion history and dark energy equation of state.

### 13.5.3 Shear cross-correlations

A very important application of lensing shear is its cross-correlation with other cosmological fields, in particular the galaxy density. So let us consider the projected galaxy density following Eq. (11.39),

$$\Delta_g(\boldsymbol{\theta}) = \int_0^\infty d\chi\, W_g(\chi)\delta_g\left(\boldsymbol{x} = \boldsymbol{\theta}\chi, \eta = \eta_0 - \chi\right), \tag{13.59}$$

where we have introduced another weighting function $W_{\rm g}(\chi)$, since this galaxy sample does not have to be the same as the source galaxies used for measuring shear. The flat-sky and Limber approximations that led to Eq. (11.47) are not restricted to auto-correlations. The result for the cross-correlation simply replaces one of the projection kernels with the appropriate one for the field that is being correlated with. So analogously to Eq. (13.46), we obtain

$$C_{{\rm g}E}(l) = l^2 \int_0^\infty \frac{d\chi}{\chi^2} W_{\rm g}(\chi) \frac{g_{\rm L}(\chi)}{\chi} P_{{\rm g},\Phi}\left(k = \frac{l+1/2}{\chi}, \eta(\chi)\right), \tag{13.60}$$

where $P_{{\rm g},\Phi}$ is the cross-power spectrum between the 3D galaxy density and the potential. Using the Poisson equation again yields

$$C_{{\rm g}E}(l) = \frac{3}{2}\Omega_{\rm m} H_0^2 \int_0^\infty \frac{d\chi}{\chi} W_{\rm g}(\chi) a^{-1}(\chi) g_{\rm L}(\chi) P_{\rm gm}\left(k = \frac{l}{\chi}, \eta(\chi)\right), \tag{13.61}$$

which involves the 3D galaxy-matter cross power spectrum. So what does this result mean? Consider the integrand at fixed $\chi$, which is the distance to the galaxies whose positions we are correlating with ("g"). The kernel $g_{\rm L}(\chi)$ is only significant if $\chi$ is significantly smaller than the typical comoving distance of the source galaxies, those used for the shear measurement. That is, there only is a nonzero cross-correlation if the source galaxies are at higher redshifts than the galaxies whose positions we are correlating with. This is because $C_{{\rm g}E}(l)$ measures the lensing effect of the *mass associated with these galaxies at lower redshifts*. So the galaxies included in $\Delta_{\rm g}$ are usually referred to as "lens galaxies," as opposed to the "source galaxies" used to measure the shear.

One of the important ramifications of Eq. (13.61) is that $P_{\rm gm}$ is proportional to $b_1$, i.e. linearly proportional to the bias of the lens galaxies, while the angular galaxy auto-correlation $C_{\rm g}(l)$ is proportional to $b_1^2$. So, by measuring both $C_{{\rm g}E}(l)$ and $C_{\rm g}(l)$ we have a means of breaking the bias-amplitude degeneracy. In fact, current lensing analyses of large imaging surveys typically obtain their tightest constraints when combining all three two-point correlations: $C_{EE}(l)$, $C_{{\rm g}E}(l)$, $C_{\rm g}(l)$ (or their real-space counterparts). Importantly however, the simple description of $P_{\rm gm}$ and $P_{\rm g}$ in terms of a linear bias $b_1$ is not expected to hold on small, nonlinear scales.

Next, let us derive the shear-galaxy cross-correlation function in real space. Again, we will align the $x$-axis with $\boldsymbol{\theta}$. So, $\gamma_1$ becomes the tangential shear $\gamma_t$. We obtain

$$\begin{aligned}\xi_{{\rm g},t}(\theta) \equiv \left\langle \Delta_{\rm g}(\mathbf{0})\gamma_t(\boldsymbol{\theta})\right\rangle &= \int \frac{d^2l}{(2\pi)^2} e^{il\theta\cos\phi_l} \cos 2\phi_l C_{{\rm g}E}(l) \\ &= -\int \frac{ldl}{2\pi} J_2(l\theta) C_{{\rm g}E}(l).\end{aligned} \tag{13.62}$$

It is easy to show that the corresponding cross-correlation with $\gamma_\times$ (i.e. $\gamma_2$ in the chosen coordinates) vanishes in the absence of $B$-modes. So $\xi_{{\rm g},\times}(\theta)$ can be used as another observational systematics check. There is an intuitive explanation for this fact. As we have seen, the galaxy-shear cross-correlation probes the lensing induced by the mass associated with

the lens galaxies. Now, there is no preferred direction on the sky, so the average mass distribution around the lens galaxies is azimuthally symmetric. So, if we place the lens galaxies at the origin, we are in a similar setup as illustrated in Fig. 10.4. This shows that $E$-modes only generate tangential shear $\gamma_t$, not the cross-component $\gamma_\times$. More precisely, Fig. 10.4 shows a radial oscillating wave, while in the case of lensing by an overdensity, the $E$-mode is always positive so that galaxy shapes will always be aligned tangentially.

Fig. 13.9 shows the measurements of $\xi_{g,t}(\theta)$ from DES. The lens galaxies are selected based on their photometric properties to be luminous red galaxies whose photometric redshifts are fairly accurate. Each panel corresponds to a different redshift range of lens galaxies (we show the auto-correlation of one of these samples in Fig. 11.9). In each panel, different sets of points show different source galaxy samples. The signal increases with increasing source galaxy redshift, and becomes small if the source galaxies are not sufficiently far behind the lenses. All of this is expected based on the shape of the lensing kernel, Eq. (13.49). While the signal-to-noise is again highest on small scales (toward the left in the figure), nonlinear bias, intrinsic alignments, and other nonlinear effects become important on those scales. For this reason, the data in the shaded regions in Fig. 13.9 are

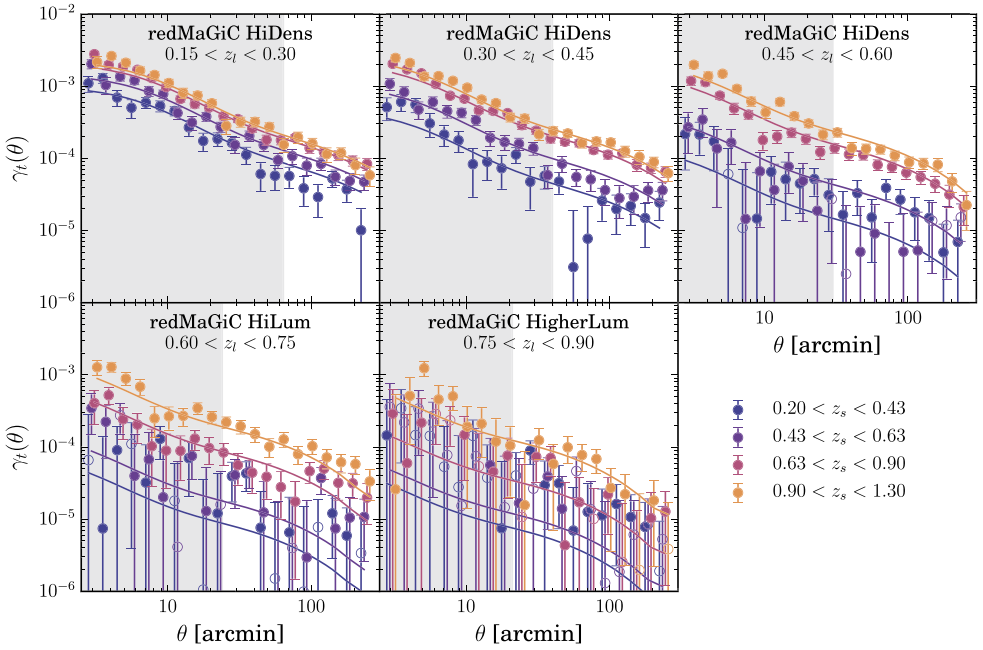



**FIGURE 13.9** Measurement of galaxy-galaxy lensing ($\gamma_t(\theta) \equiv \xi_{g,t}(\theta)$), the shear-galaxy density cross-correlation function, in the DES Year 1 data. Each panel corresponds to a different sample of lens galaxies, while the different sets of points show different source galaxy samples. The upper right panel shows results for the same lens galaxies whose density auto-correlation is shown in Fig. 11.9. Only $\xi_{g,t}(\theta)$ is shown, since $\xi_{g,\times}(\theta)$ was found to be consistent with zero, as expected. The lines show fits based on a $\Lambda$CDM cosmology, which are used to measure the bias of the lens galaxy samples. From Prat et al. (2018).

*not* used for cosmological constraints. Better models of these nonlinear scales will hopefully allow for more information to be extracted from the data in the future.

Our result Eq. (13.61) for $C_{gE}(l) = C_{g\kappa}(l)$ also allows us to understand another application of lensing, *cluster mass calibration*. $\xi_{g,t}(\theta)$ is essentially the mean mass profile around lens galaxies projected on the sky (and weighted with the lensing kernel). We can apply this measurement technique to a set of galaxy clusters selected by a mass proxy such as richness, X-ray temperature, or SZ signal. Then, lensing allows us to measure the projected mass profile of the clusters at fixed mass proxy, and hence the mean mass-observable relation.

## 13.6 Summary

Gravitational lensing is a powerful tool in that it directly probes all clustering components —dark or luminous—in the universe through their gravitational effect. We only briefly touched on the many facets of this very broad topic in Sect. 13.1, and refer the reader interested in more depth to the comprehensive review of Bartelmann and Schneider (2001). In cosmology, the most prominent application of lensing is *weak lensing*, the statistical detection of lensing through its small but ubiquitous effect on the CMB and the shapes of background galaxies. Starting from the fundamental property of weak lensing, that it conserves surface brightness or intensity (Eq. (13.3))

$$I_{\rm obs}(\boldsymbol{\theta}) = I_{\rm true}(\boldsymbol{\theta}_S), \tag{13.63}$$

where $\boldsymbol{\theta}_S(\boldsymbol{\theta})$ is the sky position in the absence of lensing, weak lensing can be completely described by a remapping of positions on the sky through

$$\begin{aligned} \theta^i_S(\boldsymbol{\theta}) &= \theta^i + \Delta\theta^i, \\ \Delta\theta^i(\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta^i}\phi_{\rm L}(\boldsymbol{\theta}); \quad \phi_{\rm L} = 2\int_0^\chi \frac{d\chi'}{\chi'}\Phi_{,i}(\boldsymbol{x}(\boldsymbol{\theta},\chi'))\left(1-\frac{\chi'}{\chi}\right), \end{aligned} \tag{13.64}$$

where $\chi$ is the comoving distance to the source. The deflection angle $\Delta\theta^i$ on the sky is the basic lensing observable. For galaxies, we most easily observe its derivatives, the *convergence* ($\kappa$) and *shear* ($\gamma$) fields, which affect the observed shapes of galaxies (Eq. (13.40)).

The approach to compare theory with data is then very similar to the other probes we have studied: one looks at statistics such as two-point correlation functions and power spectra. The effect of lensing on the CMB can be predicted (Eq. (13.21)) and probes structure at $z \simeq 2-5$. Galaxy shape correlations can be described using the same formalism we developed for polarization in Ch. 10, via a decomposition into $E$- and $B$-modes. Lensing also affects the number and size correlations of galaxies, which can be used to measure the magnification ($\mu$) instead of the shear. All of these statistics are at leading order determined by the power spectrum

$$C_{EE}(l) = C_{\kappa\kappa}(l) = \left(\frac{3}{2}\Omega_{\rm m}H_0^2\right)^2 \int_0^\infty d\chi\, a^{-2}(\chi) g_{\rm L}^2(\chi) P\left(k=\frac{l}{\chi},\eta(\chi)\right) \tag{13.65}$$

of the shear $E$-mode, which is equal to that of the convergence $\kappa = \mu/2$. The lensing kernel $g_{\rm L}(\chi)$ is defined in Eq. (13.49). The expected null result for $B$-modes on the other hand can be used as a powerful systematics check in observational measurements.

The shear power spectrum is sensitive not only to the nonlinear matter power spectrum $P(k,\eta)$ integrated along the line of sight, but also to the background cosmology through the distance-redshift relation $\chi(z)$. So, if the redshift distribution of the source galaxies is known, lensing allows for constraints on the histories of both expansion and growth. Apart from having to determine the redshift distribution of the source galaxies, intrinsic galaxy shape correlations and small-scale baryonic effects on the matter distribution itself also need to be modeled.

As powerful as the shear power spectrum itself are cross-correlations of shear with foreground galaxy number counts (galaxy-galaxy lensing) or clusters, which we derived in Eq. (13.61):

$$C_{gE}(l) = \frac{3}{2}\Omega_{\rm m} H_0^2 \int_0^\infty \frac{d\chi}{\chi} W_{\rm g}(\chi) a^{-1}(\chi) g_{\rm L}(\chi) P_{\rm gm}\left(k = \frac{l}{\chi}, \eta(\chi)\right). \tag{13.66}$$

This cross-power spectrum allows for a calibration of the bias of the lens galaxies (since $P_{\rm gm}(k) = b_1 P_{\rm L}(k)$ on large scales), and leads to a significant increase in constraining power when combined with galaxy and shear auto-correlations. For clusters, weak-lensing shear allows for the important calibration of the mass-observable relation.

## Exercises

**13.1** Derive the relation Eq. (13.2) between the specific intensity $I_\nu(\boldsymbol{x}, \hat{\boldsymbol{p}}, t)$ and the photon distribution function. Use the definition in Eq. (13.1) and $|\boldsymbol{p}| = 2\pi\nu$ (recall that we set $\hbar \equiv h/2\pi = 1$). Verify that this recovers Eq. (1.9) for a black-body spectrum.

**13.2** Derive Eq. (13.21). This is analogous to (but a bit easier than) the derivation leading from the perturbation-theory expansion of the nonlinear density field Eq. (12.40) to the next-to-leading correction to the matter power spectrum Eq. (12.48). Hence, you probably want to go through Exercise 12.5 first. Take into account the fact that $\langle\Theta(\boldsymbol{l})\phi_{\rm L}(\boldsymbol{l}')\rangle = 0$.

**13.3** Fill in some of the details leading to Eq. (13.39).

**(a)** Derive Eqs. (13.34)–(13.35), as well as the limiting expression for the magnification $\mu = |A|^{-1}$ when all of $\kappa, \gamma_1, \gamma_2$ are small.

**(b)** Derive Eqs. (13.38)–(13.39) from Eq. (13.37).

**13.4** Consider a lens at a fixed comoving distance $\chi_L$ and redshift $z_L$ (i.e., a single galaxy or a cluster as opposed to large-scale structure in general; cf. Fig. 13.4). Assume further that the extent in the line-of-sight direction is small compared to $\chi_L$. Show that the lensing potential $\phi_{\rm L}$ in Eq. (13.16) can be written as

$$\phi_{\rm L}(\boldsymbol{\theta}; \chi_L) = \frac{4G}{(1+z_L)^2}(\chi - \chi_L)\frac{\chi_L}{\chi}\int d^2\theta'\, \Sigma(\boldsymbol{\theta}') \ln\left|\boldsymbol{\theta}' - \boldsymbol{\theta}\right|. \tag{13.67}$$

Here, $\chi$ is the comoving distance out to the source and $\Sigma(\boldsymbol{\theta})$ is the projected surface density of the lens in the plane perpendicular to the line of sight:

$$\Sigma(\boldsymbol{\theta}) = \int_0^\infty d\chi'\, \rho(\boldsymbol{x}(\boldsymbol{\theta}, \chi')). \tag{13.68}$$

Hint: use cylindrical coordinates and the integral solution of the Poisson equation.

**13.5** Throughout this chapter, we considered scalar perturbations, which allowed us to write the linear-order lensing deflection as the gradient of a scalar lensing potential, Eq. (13.16). Now consider a curl contribution to $\Delta\theta$:

$$\Delta\theta_i = \partial_{\theta,i}\phi_{\rm L} + \epsilon_{3ij}\partial_{\theta,j}\omega, \tag{13.69}$$

where $\epsilon_{ijk}$ is the Levi-Civita symbol and we continue to assume that the line of sight is along the $z$-direction ($k = 3$). $\omega$ could be induced by vector and tensor metric perturbations, or by higher-order lensing contributions. Compute the contribution to $\psi_{ij}$ due to $\omega$. Show that it contributes to the shear as well as inducing a rotation (which we have dropped from $\psi_{ij}$ in the main text). Then, decompose the shear induced by $\omega$ into $E$- and $B$-modes. We thus have four components of $\psi_{ij}$: $\kappa$, $\gamma_1$, $\gamma_2$ and the rotation. Summarize the relations between the four quantities and with $\phi_{\rm L}$ and $\omega$.

**13.6** Compute $C_{\kappa\kappa}(l)$ numerically for the fiducial $\Lambda$CDM cosmology, using the linear matter power spectrum. Assume all background galaxies are at redshift $z = 1.5$. Now repeat the same calculation with a prescription for the nonlinear matter power spectrum, or compare with Fig. 13.7. At what $l$ do you begin to see a discrepancy? How would you estimate this scale $l$ based on what you have learned about matter nonlinearities in Ch. 12 and lensing in this chapter?

**13.7** Repeat Exercise 13.6 but now for a source redshift $z_* = 1089$, i.e. the CMB. This time, plot the angular power spectrum of the lensing potential $\phi_{\rm L}$, as well as that of its gradient $\Delta\theta^i = \partial\phi_{\rm L}/\partial\theta^i$. What do you conclude about the typical size of CMB lensing deflections, and their typical correlation scale?

**13.8** Derive the RMS amplitude of $\kappa$ and shear $E$-modes given an angular Fourier-space filter $W(l)$. Using the result from Exercise 13.6, evaluate it numerically for a "sharp-$l$" filter

$$W(l) = \begin{cases} 1, & l \le \pi/\theta_{\rm min}, \\ 0, & \text{otherwise}, \end{cases} \tag{13.70}$$

as a function of $\theta_{\rm min}$. This gives a sense of the typical values of the convergence and shear. What do you conclude about the size of weak lensing?

**13.9** Compute the effect of linear tidal alignments (Eq. (13.52)) on shear statistics, by following closely the derivation in Sect. 13.5.1.

**(a)** The simplest approach to incorporate alignments is by adding them to the lensing potential:

$$\phi_{\rm L}(\boldsymbol{\theta}) \to \phi_{\rm L}(\boldsymbol{\theta}) + \phi_{\rm IA}(\boldsymbol{\theta}). \tag{13.71}$$

Derive $\phi_{\rm IA}$ and its associated projection kernel $W_{\rm IA}$.

**(b)** Derive the contributions to $C_{EE}(l)$ in the Limber approximation. You should obtain two new types of contributions. Provide physical interpretations for them.

**(c)** Now follow the derivation in Sect. 13.5.3 to derive the alignment contribution to $C_{{\rm g}E}(l)$. Again provide physical interpretations for the terms you obtain.

**(d)** Can you think of approaches that could be used to disentangle lensing and intrinsic alignments?

**13.10** Derive the contribution of shear $B$-modes to the correlation functions Eqs. (13.57)–(13.58). Begin by using Eq. (10.11) to write the contribution to the components $\gamma_1(\boldsymbol{l})$, $\gamma_2(\boldsymbol{l})$.

**13.11** In Ch. 11, we computed the angular auto-correlations of galaxies, while in this chapter we derived the cross-correlation with shear. There is an additional lensing effect that we neglected in these derivations. Typically, galaxies are selected, at least to some level, by their observed flux. This means that some galaxies that should not be included in the survey because they are intrinsically fainter than the magnitude limit are magnified and so appear brighter, thereby making the cut; this is referred to as *magnification bias* (Moessner and Jain, 1998). If the magnification is $\mu$, then the number of background galaxies in an angular patch is

$$n_{\rm g} = \bar{n}_{\rm g}\mu^{2.5s-1}. \tag{13.72}$$

Here $\bar{n}_{\rm g}$ is the average number of galaxies, and $s$ is defined as $d\log N(m)/dm$ where $N(m)$ is the number of galaxies at the magnitude limit $m$. For the present exercise, do not worry about where this relation comes from (see Broadhurst et al., 1995, for an explanation).

**(a)** Derive the contribution to the observed three-dimensional galaxy density $\delta_{\rm g}$ induced by this effect at linear order in $\kappa$.

**(b)** Now derive the contribution to the projected galaxy density $\Delta_{\rm g}(\boldsymbol{\theta})$, using Eq. (13.49).

**(c)** Compute the contribution to the galaxy-shear cross-power spectrum, and cross-correlation function.

**(d)** Do the same for the cross-correlation of number counts of different galaxy samples. When do you expect the magnification bias contribution to be significant?

# 14

# Analysis and inference

Increasingly, cosmologists are turning their attention to the fundamental question of how best to analyze a set of data. The main reason for this focus is that the quality and quantity of data have improved dramatically over the past decades. There is every reason to believe that this trend will continue. Anisotropies in the temperature of the CMB have been measured by dozens of experiments and the next stage of experiments is already ramping up. The large-scale clustering of matter is probed in a variety of ways; activity here, too, shows no sign of letting up. After the completion of the Sloan Digital Sky Survey (SDSS) and the Two Degree Field (2dF) galaxy redshift survey, surveyors carried out large galaxy lensing surveys such as the Kilo-Degree Survey, the Dark Energy Survey, and Hyper Suprime-Cam survey, as well as spectroscopic follow-ups to SDSS. In the 2020s, the Dark Energy Survey Instrument (DESI), the Euclid satellite, and the Legacy Survey of Space and Time (LSST) at the Vera Rubin Observatory will dominate the landscape. The huge data sets delivered by these surveys create new challenges in the analysis.

As the data sets get larger, simple algorithms that have been traditionally used to analyze data no longer suffice. For one, the size of data sets is growing exponentially on time scales similar to computational power. Imagine that you are working on an experiment and use an algorithm that scales as $m^2$ where $m$ is the number of data points. In two years, say, your data size has doubled, so the analysis requires four times as many computing operations. Even if you now have access to a computer that is twice as fast, your analysis will still take twice as long as before. This part of the problem is exacerbated by the fact that most analyses need to be run multiple times on a given data set or simulation that mimics it. On a more profound level, as the size of data sets increases, so does their statistical precision: statistical errors on quantities typically scale as $m^{-1/2}$, as we will see. As statistical errors go down, all kinds of issues that were previously buried under large error bars and were therefore never considered need to be accounted for. These are so-called systematic errors, which do not fall off as $m^{-1/2}$. Importantly, that applies to both instrumental artifacts as well as physical effects that were not modeled properly in the theory prediction. We will not go into specific systematic issues in this chapter, since they are highly specific to individual instruments and techniques.

The analysis techniques we focus on in this chapter are tools developed in order to deal with the complexity of current cosmological data, but also are broadly applicable to many other research areas that involve large data sets. We begin with an introduction to the concepts of likelihood, prior, and posterior in the context of a simple example in Sect. 14.1, which any reader not already familiar with these terms should read first. Sect. 14.2 provides an overview of one of the most common ways these concepts are applied to cosmology: to constrain cosmological parameters via a measured power spectrum of some form. The rest

**Modern Cosmology.** https://doi.org/10.1016/B978-0-12-815948-4.00020-6
Copyright © 2021 Elsevier Inc. All rights reserved.

of the chapter fleshes out the steps in this prototypical example, depicted in Fig. 14.1. In particular, in Sect. 14.3 and Sect. 14.4, we discuss how to construct estimators of power spectra as well as their likelihood.

Neglecting all further experimental issues and systematics, this is in principle all we need to infer cosmological parameters from the power spectra. However, real-world issues including the above-mentioned size of the data sets add additional stumbling blocks. Thus, we will introduce some crucial techniques to get around these obstacles: the Fisher matrix (Sect. 14.5) is a shortcut to obtain approximate estimates of expected error bars on parameters, while Markov Chain Monte Carlo (MCMC) sampling (Sect. 14.6) allows for an efficient determination of best-fit parameters and error bars even for very complex likelihoods. Both of these techniques are applicable much more broadly than just in cosmology or astrophysics.

## 14.1 The likelihood function

The basic building block of contemporary analyses is the likelihood function. This is defined as the probability that an experiment yields the observed data given a theory. This seemingly simple definition is exceedingly powerful. Once we have the likelihood function, we can determine the parameters of the theory along with errors. Let us study this with a simple example.

Suppose you want to weigh somebody. Since you are a scientist, you know that, in addition to the measurement, you should also report an uncertainty. So you set up 100 different scales and record the person's weight on each of these different scales. Given these 100 numbers, what value should you report for the weight and the uncertainty in the weight? Let us then introduce the formalism of the likelihood function $\mathcal{L}$ in this simple context.

The likelihood function gives the probability of getting the hundred numbers given a theory. Our theory will be that each measurement is the sum of a constant signal $w$ (the person's weight) and noise, with the noise drawn from a Gaussian distribution with mean zero and variance $\sigma_w^2$. Thus our "theory" has two free parameters $w$ and $\sigma_w$. If only one data point $d$ was taken, the likelihood $\mathcal{L}$, i.e. the probability of getting $d$ given the theory would be

$$\mathcal{L}(d|w,\sigma_w) \equiv P(d|w,\sigma_w) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{(d-w)^2}{2\sigma_w^2}\right\}. \tag{14.1}$$

Here and throughout, $P(x|y)$ denotes the probability of $x$ given $y$. Eq. (14.1) restates the assumptions that $d-w$ is only sourced by noise and that the noise is drawn from a Gaussian distribution with standard deviation $\sigma_w$. In the limit that $\sigma_w$ becomes very small, this function becomes sharply peaked at $d=w$. Since we are making $m=100$ independent measurements, the likelihood function is the product of all the individual likelihood functions. That is,

$$\mathcal{L}\left(\{d_i\}_{i=1}^m \middle| w,\sigma_w\right) = \frac{1}{(2\pi\sigma_w^2)^{m/2}} \exp\left\{-\frac{\sum_{i=1}^m (d_i-w)^2}{2\sigma_w^2}\right\}. \tag{14.2}$$

Notice that, although the data are drawn from a Gaussian, the likelihood function is not Gaussian in all the theoretical parameters (it is in $w$ but is not in $\sigma_w$).

We are interested in the value of the theoretical parameters $w$ and $\sigma_w$. Thus, instead of $P(\{d_i\}|w,\sigma_w)$, which is the likelihood function we have written down, we want $P(w,\sigma_w|\{d_i\})$. To obtain the latter from the former we can use a relation from probability theory,

$$\begin{aligned} P(B,A) &= P(B|A)P(A) \\ &= P(A|B)P(B). \end{aligned} \tag{14.3}$$

In this context, $A = \{d_i\}$ is the set of measurements, and $B = \{w,\sigma_w\}$ is the set of "model parameters," so the equality between the two lines of Eq. (14.3) means that

$$P(w,\sigma_w|\{d_i\}) = \frac{P(\{d_i\}|w,\sigma_w)\ P(w,\sigma_w)}{P(\{d_i\})}. \tag{14.4}$$

This is known as *Bayes' theorem*. The denominator is independent of the parameters $w$, $\sigma_w$, and so can be determined by realizing that when we integrate the probability $P(w,\sigma_w|\{d_i\})$ over all values of the parameters $w$, $\sigma_w$ (keeping the data fixed), we must get 1. So the denominator is equal to the integral of the numerator over $w$, $\sigma_w$. A constant normalization factor does not affect the place in parameter space where the likelihood function peaks or the width of the likelihood function. For the most part, then, we are free to ignore it.

To get the probability of the "theory" given the data $P(w,\sigma_w|\{d_i\})$, which is what we want, we need the likelihood function—the first term in the numerator—and the *prior* probability $P(w,\sigma_w)$. If we possess prior information about these theory parameters, we should use this information here. Then,

$$P(w,\sigma_w|\{d_i\}) \propto \mathcal{L}\left(\{d_i\}_{i=1}^m \middle| w,\sigma_w\right) P_{\text{prior}}(w,\sigma_w), \tag{14.5}$$

the proportionality constant being independent of the parameters and therefore of little interest. The resulting probability distribution is called the *posterior* for $w$, $\sigma_w$ given the data. As cosmologists, we are most interested in the posteriors for cosmological parameters such as the dark energy equation of state or the tensor-to-scalar ratio.

The idea of using prior information might seem unsatisfying, since it appears to introduce an ambiguity into the posterior. If we want to be conservative, and assume nothing, we put in a uniform prior for the parameters, but even this choice is not as innocent as it sounds. If we had taken the parameter to be $\sigma_w^2$ instead of $\sigma_w$ and we had assumed a prior uniform in $\sigma_w^2$, i.e., that equal intervals of $\sigma_w^2$ are equally likely, we would get a different answer for the final posterior (try it!). The primary purpose of the prior is to allow us to include additional information. For example, if the manufacturer of the scale in our example told you that $\sigma_w$ was lower than some value, then the prior would permit you to incorporate that information in a consistent way. There are a wide variety of (in)famous examples wherein accounting for the prior is of tremendous importance (see Exercise 14.1).

We will now adopt a uniform prior for the parameters $w$ and $\sigma_w$ and can then find best-fit values for them. Simply find the place in parameter space where $P(w, \sigma_w|\{d_i\})$ is largest. In this simple example, we can proceed analytically by differentiating $\mathcal{L}$ with respect to each of the parameters. First consider the derivative with respect to $w$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\sum_{j=1}^{m}(d_j - w)}{\sigma_w^2 (2\pi\sigma_w^2)^{m/2}} \exp\left\{-\frac{\sum_{i=1}^{m}(d_i - w)^2}{2\sigma_w^2}\right\}. \tag{14.6}$$

For this derivative to be zero, we have

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \qquad \Leftrightarrow \qquad \sum_{j=1}^{m}(d_j - w) = 0 \tag{14.7}$$

or equivalently, the likelihood is at a maximum when

$$w = \hat{w} = \frac{1}{m}\sum_{i=1}^{m} d_i. \tag{14.8}$$

This is how we would estimate the weight from the data, and appropriately, $\hat{w}$ is called an *estimator* for the weight; in this case, it is simply the sample mean. In Exercise 14.2, you will derive the maximum-likelihood value if each data point has a different error $\sigma_{w,i}$, resulting in *inverse-variance* weighting.

The variance $\sigma_w^2$ is also a parameter in the model. So, we can find what the most probable value of $\sigma_w^2$ is given the data by computing

$$\frac{\partial \mathcal{L}}{\partial \sigma_w^2} = \mathcal{L} \times \left[-\frac{m}{2\sigma_w^2} + \frac{\sum_{i=1}^{m}(d_i - w)^2}{2\sigma_w^4}\right] \tag{14.9}$$

and setting it equal to zero. Solving for the variance $\sigma_w^2$, we find a most probable value of

$$\widehat{\sigma_w^2} = \frac{1}{m}\sum_{i=1}^{m}(d_i - w)^2. \tag{14.10}$$

This serves as an estimator for the variance from the data, for a known mean $w$.[1]

We have found estimators for our theoretical parameters. What is the error on these estimated values? The error essentially corresponds to the width of the likelihood function. More rigorously, the posterior $P(w|\{d_i\})$, i.e. the likelihood multiplied by the prior, allows us to construct confidence intervals. For example, the values of $w$ on either side of the maximum—call them $w_-$ and $w_+$—that $(i)$ have the same probability and $(ii)$ satisfy

$$\int_{w_-}^{w_+} dw\, P(w|\{d_i\}) = 0.68 \tag{14.11}$$

[1] Accounting for the fact that $w$ is estimated from the same data changes the normalization from $1/m$ to $1/(m-1)$, yielding the standard sample variance.

define the 1-$\sigma$ (or 68% confidence-level) error bar; that is, the lower 1-$\sigma$ error is the difference between the value at which the probability peaks and $w_-$. The value 0.68 stems from the Gaussian probability to find a value within 1-$\sigma$ of the mean. In the case of Eq. (14.2), this rigorous definition translates into the condition that, at the boundaries of the 1-$\sigma$ confidence interval, the log-likelihood is reduced from the maximum value by $\Delta \ln \mathcal{L} = -1/2$.

In our example, where the likelihood is Gaussian in $w$, the errors are symmetric and we can easily compute them. Expanding the log of the likelihood about the maximum for $w$, we have

$$\begin{aligned}\ln \mathcal{L}(w) &= \ln \mathcal{L}(\hat{w}) + \left.\frac{\partial \ln \mathcal{L}}{\partial w}\right|_{w=\hat{w}} (w-\hat{w}) + \frac{1}{2}\left.\frac{\partial^2 \ln \mathcal{L}}{\partial w^2}\right|_{w=\hat{w}} (w-\hat{w})^2 \\ &= \ln \mathcal{L}(\hat{w}) - \frac{m}{2\sigma_w^2}(w-\hat{w})^2, \end{aligned} \tag{14.12}$$

since $\partial \ln \mathcal{L}/\partial w$ vanishes at $\hat{w}$. Equating this to $\ln \mathcal{L}(\hat{w}) - 1/2$, the condition for 1-$\sigma$ confidence intervals, we have

$$\mathrm{Var}\left[\hat{w}\right] = \frac{\sigma_w^2}{m}. \tag{14.13}$$

The square root of this variance, $\sigma_w/m^{1/2}$, is the $1\sigma$ error in our estimate of $w$. This too is familiar: as more measurements are taken, the noise gets beaten down by a factor of the square root of the number of independent measurements. More generally, this derivation leads to the rule that the variance of a one-dimensional Gaussian distribution is proportional to the inverse of the coefficient of the quadratic term in the log-likelihood.

Above, we computed the uncertainty on our estimate of $w$ by considering the width of the likelihood function. There is another way to obtain this uncertainty, one that generalizes to more complex problems, so let us use this approach here to determine the variance of the estimator for $\sigma_w^2$. It is defined as

$$\mathrm{Var}(\sigma_w^2) \equiv \left\langle \left(\widehat{\sigma_w^2} - \sigma_w^2\right)^2 \right\rangle. \tag{14.14}$$

Here, the angular brackets denote the expected value if one were to perform the experiment an infinite number of times. We can calculate this variance by integrating it over all possible values of the data weighted by the likelihood. So, in general, for any estimator $\hat{O}$, i.e. any function of the $m$ measurements $\{d_i\}$,

$$\left\langle \hat{O} \right\rangle = \int d(d_1) \int d(d_2) \ldots \int d(d_m)\, \hat{O}(\{d_i\}) \mathcal{L}(\{d_i\}). \tag{14.15}$$

To be clear, the expression for the variance contains $\widehat{\sigma_w^2}(\{d_i\})$ and $\sigma_w^2$; the latter is a number, the true value that quantifies the precision of the scales. The former is an estimator whose

expectation value is equal to the number: $\left\langle \widehat{\sigma_w^2} \right\rangle = \sigma_w^2$. This simplifies the calculation:

$$\begin{aligned} \mathrm{Var}(\widehat{\sigma_w^2}) &= \left\langle (\widehat{\sigma_w^2})^2 \right\rangle - 2\left\langle \widehat{\sigma_w^2} \right\rangle \sigma_w^2 + \sigma_w^4 \\ &= \left\langle (\widehat{\sigma_w^2})^2 \right\rangle - \sigma_w^4. \end{aligned} \tag{14.16}$$

The calculation then hinges on $\left\langle (\widehat{\sigma_w^2})^2 \right\rangle$. For simplicity, let us assume that the mean is known so we can shift all the integration variables in Eq. (14.15) by $d_i \to x_i \equiv d_i - w$. Then, we can do the calculation as if the mean were zero:

$$\left\langle (\widehat{\sigma_w^2})^2 \right\rangle = \frac{1}{m^2} \sum_{ij} \left\langle x_i^2 x_j^2 \right\rangle. \tag{14.17}$$

Picking a term with $i \neq j$ from the sum, we have

$$\begin{aligned} \left\langle x_i^2 x_j^2 \right\rangle &= \left[ \prod_{k \neq i,j} \int dx_k \, \frac{e^{-x_k^2/2\sigma_w^2}}{\sqrt{2\pi\sigma_w^2}} \right] \left[ \int dx_i \, x_i^2 \frac{e^{-x_i^2/2\sigma_w^2}}{\sqrt{2\pi\sigma_w^2}} \right] \left[ \int dx_j \, x_j^2 \frac{e^{-x_j^2/2\sigma_w^2}}{\sqrt{2\pi\sigma_w^2}} \right] \\ &= \left[ \int dx_i \, x_i^2 \frac{e^{-x_i^2/2\sigma_w^2}}{\sqrt{2\pi\sigma_w^2}} \right] \left[ \int dx_j \, x_j^2 \frac{e^{-x_j^2/2\sigma_w^2}}{\sqrt{2\pi\sigma_w^2}} \right] \end{aligned} \tag{14.18}$$

since all the integrals over $x_k$ are equal to 1. In the second line, each integral is equal to $\sigma_w^2$; the sum over $j \neq i$ in Eq. (14.17) yields a factor of $m - 1$, and then the sum over $i$ another factor of $m$, so

$$\mathrm{Var}(\widehat{\sigma_w^2}) = \frac{m-1}{m} \sigma_w^4 + \frac{1}{m^2} \sum_i \left\langle x_i^4 \right\rangle - \sigma_w^4. \tag{14.19}$$

By using Eq. (14.15) and integrating by parts (or using Wick's theorem), we see that $\langle x_i^4 \rangle = 3\langle x_i^2 \rangle^2 = 3\sigma_w^4$, so

$$\mathrm{Var}(\widehat{\sigma_w^2}) = \frac{2}{m} \sigma_w^4. \tag{14.20}$$

Equivalently, the uncertainty on the variance is

$$\sqrt{\mathrm{Var}\left[\widehat{\sigma_w^2}\right]} = \sqrt{\frac{2}{m}} \sigma_w^2. \tag{14.21}$$

This error on $\sigma_w$ may seem a rather arcane fact, but it turns out that most of what we are interested in measuring in the realm of cosmology is akin to $\sigma_w^2$. The fluctuations around the mean galaxy density and CMB temperature are of much more interest than the mean quantities themselves. These fluctuations are drawn from distributions—often approximately Gaussian—the parameters of which we are keenly interested in, because they depend on the cosmological model. Therefore, we want to know how accurately we

can measure parameters like $\sigma_w^2$, and it turns out that Eq. (14.21) is generic. When estimating the variance of a distribution, there is a fundamental uncertainty proportional to that variance divided by the square root of the number of measurements; it is called *sample variance* or *cosmic variance*.

In real applications, we typically have not only a single unknown parameter like $\sigma_w$, but several of them. If we are not interested in some parameters, call them "nuisance" parameters, that describe observational or astrophysical effects for example, then we should *marginalize* over them, i.e. work with the likelihood after integrating over these uninteresting parameters. In the context of the weight example, we can imagine the case where we are interested in measuring $w$, but do not have knowledge of $\sigma_w$. Then, given the full posterior $P(w, \sigma_w|\{d_i\})$, the desired *marginalized posterior* is

$$P(w|\{d_i\}) = \int_0^\infty d\sigma_w \, P(w, \sigma_w|\{d_i\}). \tag{14.22}$$

The left-hand side can then be used to give confidence intervals for $w$ that properly take into account our lack of knowledge of $\sigma_w$. We will see a more concrete example of this in Sect. 14.5.

## 14.2 Overview: from raw data to parameter constraints

Moving away from the weight metaphor, we can apply what we have learned about the likelihood and posterior to the CMB and other cosmic probes such as the galaxy distribution and weak lensing.

Fig. 14.1 provides an overview of these steps in the case of two-point functions, starting from the upper left and ending in a contour plot that describes parameter constraints, as sketched at the right. Each of the boxes and arrows in Fig. 14.1 represents an enormous amount of work for even one experiment, and the following sections attempt to give a flavor of the work required to carry these out. The first data product is a map: for the CMB, this could be a map of the temperature anisotropy on the sky; for a galaxy survey, a map of the three-dimensional galaxy density field; for a gravitational lensing survey, a map of galaxy ellipticities, etc. The simplest statistic we measure from this map is the two-point function (correlation function or power spectrum). The likelihood combines this observable with theoretical predictions for a given set of parameters, and the covariance, to output a single number. The likelihood is also the step where we combine multiple probes. In order to find the peak of the posterior and confidence contours for the parameters, a sampler is employed (bottom right), which computes the posterior for many, often millions of sets of parameters, or samples. Alternatively, the Fisher forecast provides an analytic shortcut to a rough approximation of the expected error bars.

In this chapter, we will restrict our treatment to two-point functions and a Gaussian likelihood, so that, using the angular power spectrum $C(l)$ e.g. of CMB anisotropies as an

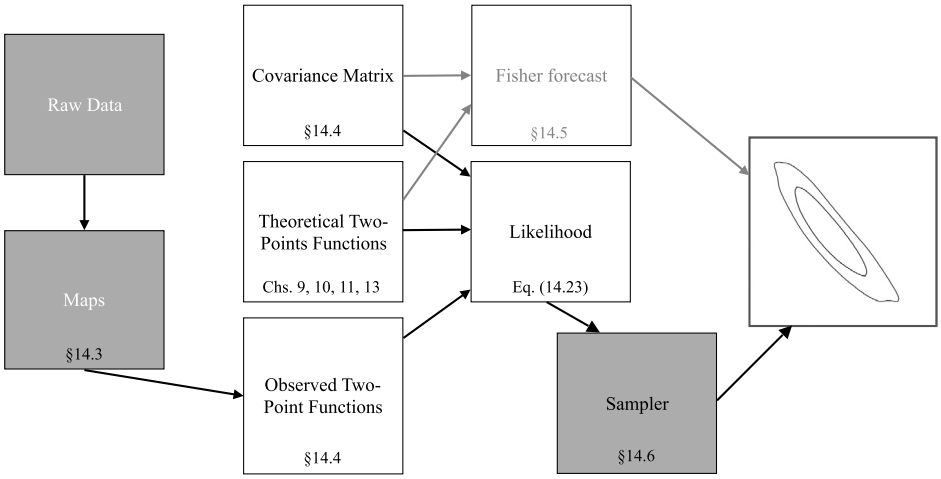



**FIGURE 14.1** Overview of how to get from raw data (upper left) to parameter constraints (right). The raw data is compressed into a map, from which two-point functions are estimated. These "Observed Two-Point Functions" are combined with a covariance matrix and the model (theory) predictions for the two-point functions, given a set of cosmological parameters, to form a likelihood for any given parameter set. A sampler calls this likelihood, multiplied in general by priors on the parameters, for many different values of parameters in order to find the preferred region in parameter space, illustrated here by the contour plot on the right. The Fisher forecast instead uses just the theoretical two-point functions and covariance to compute approximate expected likelihood contours.

example, the form of the likelihood is[2]

$$\ln \mathcal{L}(\lambda_\alpha) = -\frac{1}{2}\sum_{ll'}\left(\hat{C}(l) - C^{\rm theory}(l,\lambda_\alpha)\right)\left(\mathrm{Cov}^{-1}\right)_{ll'}\left(\hat{C}(l') - C^{\rm theory}(l',\lambda_\alpha)\right). \qquad (14.23)$$

The likelihood is computed in this case by contracting the difference between the observed (or estimated, $\hat{C}(l)$) and theoretical two-point functions ($C^{\rm theory}(l,\lambda_\alpha)$, which are a function of cosmological parameters $\lambda_\alpha$), with the inverse of the covariance matrix to obtain a single number for each set of parameters $\lambda_\alpha$.

At this point, we should note that there are other observables in cosmology than two-point functions, for example cluster counts, and a Gaussian likelihood is usually only an approximation that needs to be justified. Still, much of the action in cosmology takes place within these confines, and the tools we will develop are readily applicable to other observables and likelihood forms.

Notice that the likelihood in Eq. (14.23) is Gaussian in the observable, $\hat{C}(l)$; it is *not* Gaussian in the parameters $\lambda_\alpha$, because $C^{\rm theory}(l,\lambda_\alpha)$ is in general a very complicated function of $\lambda_\alpha$ (see for example Fig. 9.17 in the case of the CMB). Parameter degeneracies, some of which we have discussed in Sect. 9.7.2 for the CMB, add to this complexity, so that the resulting posterior contours are not nearly as simple to obtain as one might have imagined

[2] Here, we have dropped the logarithm of the determinant of the covariance, under the assumption that it is independent of the cosmological parameters $\lambda_\alpha$.

given the simple form of Eq. (14.23). Moreover, the theoretical prediction often involves free nuisance parameters that are not cosmological. One prominent example is the bias parameter $b_1$ which enters the galaxy power spectrum. We then need to map out the posterior for the cosmological parameters after integrating (marginalizing) over the nuisance parameters.

The rest of this chapter walks through the steps that are depicted in Fig. 14.1. We begin with mapmaking in Sect. 14.3. Sections 14.4.1–14.4.2 describe how to estimate $\hat{C}(l)$ and its counterpart for the galaxy power spectrum. Along the way, we will also derive expressions for the corresponding covariance matrices, which involve both cosmic variance and instrumental or astrophysical noise. Fortunately, we have already spent many pages in this book on the derivation of the theory predictions to insert into Eq. (14.23), so after Sect. 14.4 we will be all set to evaluate the likelihood. We describe sampling approaches that efficiently map out the likelihood (more precisely, posterior) in Sect. 14.6. Before that though, we detour slightly to introduce the Fisher matrix in Sect. 14.5, a handy tool, also for theorists, with which to project errors on parameters even before an experiment is done.

## 14.3 Mapmaking

The first step in the analysis of cosmological data usually consists of turning the raw data into a map of some form: for example, the CMB temperature on the sky, or the three-dimensional galaxy density field for redshift surveys. Very generically in astronomy, the raw data obtained is the sum of signal and noise. Our goal then is to combine the raw data, which we call $d_t$, into a map of the signal $s_i$ in such a way as to minimize the noise in the final map. Notice that the map has different dimensions from the data: the data might sample the signal in a given pixel $i$ many times. If the signal is not varying with time (which is the case for most cosmological data sets), then the true signal in a given pixel is a single number $s_i$. The data, however, could be tens to thousands of observations $d_t$ of that single pixel, each with their respective noise $\eta_t$. All of this can be incorporated into one equation:

$$d_t = \sum_i P_{ti} s_i + \eta_t . \tag{14.24}$$

Although Eq. (14.24) is extremely general, it is helpful to think of a specific example where a single instrument is recording the CMB flux at many times. In that case, $t$ simply labels the time of the observation. The noise in each measurement is denoted $\eta_t$; this is what we would like to remove. The index $i$ labels pixels on the sky, or in the general case, the different signals we want to estimate from the data. The matrix $P_{ti}$, in some applications called *pointing matrix*, relates the signal to the data. It is an $m_t \times m_p$ matrix where $m_t$ is the number of measurements and $m_p$ is the number of pixels (more generally signals). In perhaps the simplest of all cases, the detector at time $t$ would be taking data from one and only one pixel, so each row of $P_{ti}$ would have only a single nonzero value in it, in the column that corresponds to the pixel under observation at that time. Determining the value of that single nonzero element is called *calibration*: translating the measurement in the detector

to a flux in a pixel. Eq. (14.24) can account for any instrumental and atmospheric effects, as long as they linearly relate the signal to the data. It is even easy to visualize moving off of this simple example: if a given observation captured photons from multiple pixels, then there would be several nonzero values in a row. If there were multiple detectors, then $t$ would not just label time but rather both detector and time. You can probably think of even more complex situations, many of which are captured by this simple formula.

The noise $\eta_t$ in the data is often assumed to be Gaussian with mean zero and a covariance matrix $N_{tt'}$; we will do so here as well. There are techniques to determine $N_{tt'}$ directly from the data, but to simplify the discussion, we will assume that $N_{tt'}$ is known.

In order to derive the best way to extract the signal from the data, we consider the likelihood, more precisely its logarithm:

$$\chi^2 \equiv -2 \ln \mathcal{L}(\{d_t\}|\{s_k\}) = \sum_{tt'kl} (d_t - P_{tk} s_k) \left(N^{-1}\right)_{tt'} (d_{t'} - P_{t'l} s_l). \tag{14.25}$$

Maximizing the likelihood then is equivalent to minimizing $\chi^2$ with respect to the signal $s$. Taking the derivative of $\chi^2$ with respect to $s_i$ leads to

$$\frac{\partial \chi^2}{\partial s_i} = -2 \sum_{tt'j} P_{ti} \left(N^{-1}\right)_{tt'} (d_{t'} - P_{t'j} s_j). \tag{14.26}$$

Now we set the derivative to zero, which yields

$$\sum_{tt'j} P_{ti} \left(N^{-1}\right)_{tt'} P_{t'j} s_j = \sum_{tt'j} P_{ti} \left(N^{-1}\right)_{tt'} d_{t'}. \tag{14.27}$$

The terms multiplying $s_j$ on the left are the elements of an $m_p \times m_p$ matrix,

$$(C_N^{-1})_{ij} \equiv \sum_{tt'} P_{ti} \left(N^{-1}\right)_{tt'} P_{t'j}. \tag{14.28}$$

Multiply both sides by the inverse of this ($C_N$ itself) to find that the $\chi^2$ is minimized when $s$ is equal to

$$\hat{s}_i = \sum_{tt'j} (C_N)_{ij} \, P_{tj} \left(N^{-1}\right)_{tt'} d_{t'}. \tag{14.29}$$

In matrix notation, this is

$$\hat{s} = C_N P^\top N^{-1} d \tag{14.30}$$

where $^\top$ denotes transpose. The covariance matrix of this estimator of the signal is equal to

$$C_N = (P^\top N^{-1} P)^{-1}, \tag{14.31}$$

a fact that you can verify by taking $\langle \hat{s}_i \hat{s}_j \rangle - \langle \hat{s}_i \rangle \langle \hat{s}_j \rangle$ (Exercise 14.3).

A simple limit of Eq. (14.30) emerges when the noise matrix $N_{tt'}$ is diagonal and uniform; that is, all elements on the diagonal are identical and given by $\mathcal{N}$. In that case, the elements of $C_N$ become

$$C_{N,ij} \to \mathcal{N}\left(\sum_t P_{ti} P_{tj}\right)^{-1} \tag{14.32}$$

Assume that we only observe a single pixel at a given time, so that, for a given $t$, $P_{ti}$ is nonzero for only one pixel $i$. So the product $P_{ti} P_{tj}$ vanishes unless $i = j$ and the detector at time $t$ was pointing at pixel $i$. Thus the sum over $t$ counts the number of times the detector sampled pixel $i$; call this number $m_i$. In this simple case of uniform, uncorrelated noise, therefore, the noise covariance matrix $C_N$ for the signal estimator is diagonal with elements $\mathcal{N}/m_i$. This makes sense: as a given pixel is sampled more times, the standard deviation goes down as $m_i^{-1/2}$. The estimator for the signal now becomes

$$\hat{s}_i = \frac{1}{m_i} \sum_t P_{ti} d_t. \tag{14.33}$$

That is, one simply averages all the data points corresponding to the given pixel (exactly as in Eq. (14.8)).

Fig. 14.2 shows the result of an actual implementation of Eq. (14.30), from the Atacama Cosmology Telescope collaboration operating in Chile (Louis et al., 2017). In this case, the set of $\{\hat{s}_i\}$ is a map of the CMB temperature in many pixels on the sky. The map covers a region of 45 square degrees culled from seven months of data taking. Some additional filtering was needed to construct this map, but the basis of it is indeed Eq. (14.30).

We have tried to keep this discussion more general than just CMB mapmaking, however, because the estimator in Eq. (14.30) is broadly applicable. The only assumption we made was that there is a linear relation, with additive approximately Gaussian noise, between the data $\{d_t\}$ and the signal $\{s_i\}$ mediated by a matrix $P_{ti}$. As long as such a linear relation and the assumption of Gaussian noise hold, Eq. (14.30) is the proper estimator to use to infer the signal, in this case the map. Hence, what we have discussed so far immediately carries over to the analysis of projected galaxy clustering, i.e. to construct a map of the projected galaxy density $\Delta_g(\hat{\boldsymbol{n}})$ on the sky, or to construct a map of the lensing potential (the projected gravitational potential) from galaxy shapes (Sect. 13.4). Eq. (14.30) can be directly applied to those cases.

For galaxy redshift surveys, we first need to generalize the concept of pixels to 3D. Usually, one constructs a regular cubic grid in comoving coordinates that covers the entire survey, so that the pixels now correspond to the cells of the 3D grid. The optimal weighting encoded in Eq. (14.30) is incorporated at the level of the image and spectral analysis. Then, every galaxy is assigned to a pixel, so each pixel $i$ contains $m_{g,i}$ galaxies. The galaxy overdensity field is then defined as

$$\delta_{g,i} = \frac{m_{g,i} - \bar{m}_{g,i}}{\bar{m}_{g,i}} \tag{14.34}$$

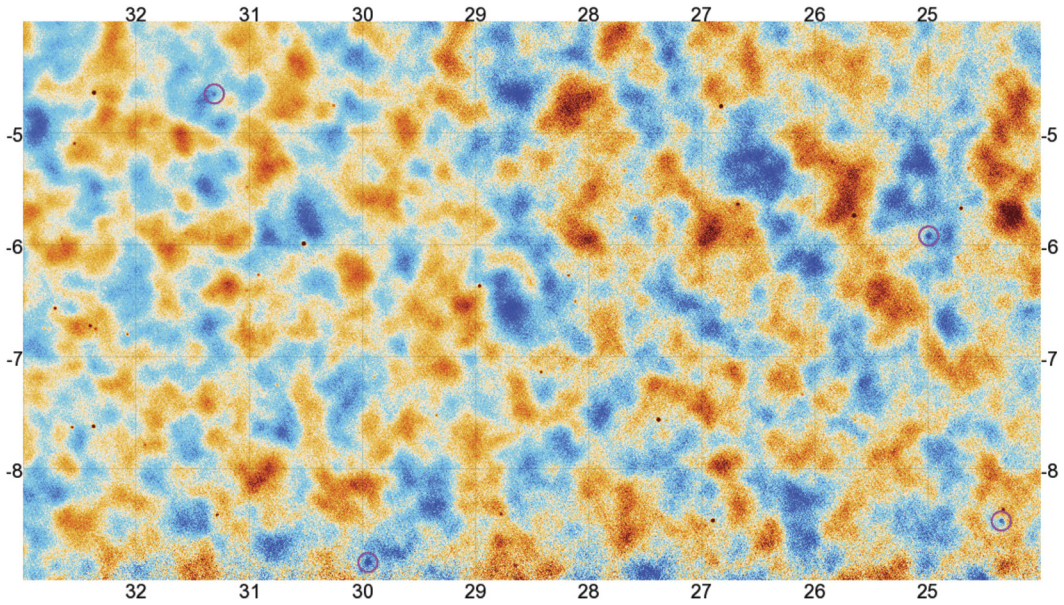



**FIGURE 14.2** A map of the CMB temperature from observations by the Atacama Cosmology Telescope. This 45 square degree region represents only a small part of their total coverage. Notice that the map has lower noise in the upper left as this is a region that was observed more frequently. Several foreground point sources are circled. From Louis et al. (2017).

where $\bar{m}_{\mathrm{g},i}$ is the number of galaxies expected in pixel $i$ if the galaxy distribution was intrinsically completely uniform. Herein lies the complexity, as this number depends on the redshift-dependent mean galaxy density. Moreover, Eq. (14.34) does not work if $\bar{m}_{\mathrm{g},i} = 0$, as is the case for pixels in the grid that lie outside the actual surveyed area, or that are "masked," for example due to a bright foreground star. We will discuss how such pixels are handled in Sect. 14.4.2.

The estimators Eq. (14.30) and Eq. (14.34) yield maps of the CMB temperature and galaxy distribution which we can then process to obtain power spectra to compare with the theory predictions, as we will describe next.

## 14.4 Two-point functions

Given a map of the CMB (or projected galaxy density field), we now need to find a way to determine the $\hat{C}(l)$ to plug into the likelihood Eq. (14.23). We have learned all the essentials we need to construct an estimator $\hat{C}(l)$ for the angular power spectrum in Sect. 14.1, and we will put them to use shortly. The main new issue we face is that our measurement process is not quite as simple as reading off someone's weight from a scale, and we will have to incorporate some observational effects such as the finite resolution of the instrument. This leads to a perennial question perhaps all scientists face when dealt a hand of data: does

one process the data so that it can be cleanly compared to simple theoretical predictions, or does one take the data as is and instead *forward-model* the theoretical prediction to account for all the observational complexities?

The second approach is in principle the right approach to take, as it cleanly separates data from interpretation (such as a fiducial cosmological model assumed in the analysis). The results of the first analysis option on the other hand can be compared not only against "simple" theory but also among different experiments. We will pursue the first option here, as it allows us to directly use the theoretical predictions we have developed in previous chapters. That is, the $C^{\rm theory}(l, \lambda_\alpha)$ in Eq. (14.23) indeed are the $C(l)$ we computed in Ch. 9 in case of the CMB temperature. The theory box in Fig. 14.1 then becomes straightforward, so we focus our attention in this section on *estimators* for the two-point functions from the maps, the box labeled "Observed Two-Point Functions" in the figure. To keep things simple, we will focus on just a single experimental complication in the case of estimates of the $C(l)$ in the CMB: the finite resolution—beam or point-spread function—of the instrument. In the galaxy clustering case, we will consider the 3D power spectrum, where the main complication is the window function of the survey.

### 14.4.1 CMB power spectrum

Let us assume we are given a map of the observed temperature anisotropies on the sky $\Delta(\hat{\boldsymbol{n}})$, inferred for instance as signal $\{s_i\}$ in pixels using the estimator derived in Sect. 14.3. We imagine that we choose the pixels small enough so that the effective smoothing induced by pixelization can be ignored, and we can treat the temperature as a continuous field. We then decompose the observed map into spherical harmonics:

$$a_{lm}^{\rm obs} = \int d\Omega\, Y_{lm}^*(\hat{\boldsymbol{n}})\Delta(\hat{\boldsymbol{n}}). \tag{14.35}$$

The superscript "obs" on $a_{lm}$ indicates that this is the observed quantity, smeared by the beam of the experiment and processed in other ways. For simplicity, let us focus just on the effect of the beam. In radio astronomy terminology, the beam describes the smearing due to the finite angular resolution of the instrument; in optical telescopes, the term point-spread function is used. Both are essentially equivalent. Including the beam smearing, the reported fractional temperature fluctuation $\Delta = (T - T_0)/T_0$ in a pixel at sky location $\hat{\boldsymbol{n}}$ is

$$\Delta(\hat{\boldsymbol{n}}) = \int d\Omega' \Theta(\hat{\boldsymbol{n}}') B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') + \eta(\hat{\boldsymbol{n}}) \tag{14.36}$$

where $B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}')$ is the beam pattern at the position $\hat{\boldsymbol{n}}$ and $\Theta$ is the true underlying temperature perturbation, while $\eta(\hat{\boldsymbol{n}})$ is the noise in the map at that position. As an example, the beam pattern of the Planck instrument at 30 GHz is shown in Fig. 14.3, that is, the figure shows the beam pattern around a fixed location $\hat{\boldsymbol{n}}$ (marked by a star) as a function of $\hat{\boldsymbol{n}}'$.

Inserting Eq. (14.36) into Eq. (14.35), we obtain

$$a_{lm}^{\rm obs} = \sum_{l'm'} B_{lm,l'm'} a_{l'm'} + \eta_{lm}, \tag{14.37}$$

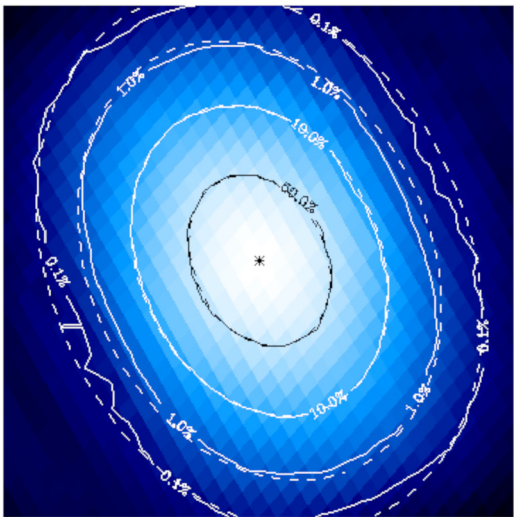



**FIGURE 14.3** The beam pattern for the Planck instrument at 30 GHz. The contours delineate the regions where the beam function drops from its maximum to 50%, 10%,1%, and 0.1%, respectively. The 50% drop occurs at roughly $30'$. Source: Planck wiki (see also Aghanim et al., 2014).

where $a_{l'm'}$ are the true CMB multipole moments (those that would be observed in the absence of beam smearing and noise), and $\eta_{lm}$ are the multipole moments of the noise. $B_{lm,l'm'}$ denotes the multipole expansion of the beam pattern in its two arguments $\hat{\boldsymbol{n}}$, $\hat{\boldsymbol{n}}'$. The second term is immediately obvious from the fact that $\eta(\hat{\boldsymbol{n}})$ simply adds to $\Delta(\hat{\boldsymbol{n}})$. You can derive the effect of the beam in Exercise 14.4.

Eq. (14.37) is very general and holds for anisotropic and spatially varying beams. In many cases, a reasonable first-order approximation is to assume the beam is constant on the sky and isotropic. You can show (Exercise 14.4) that this greatly simplifies the effect of the beam on $a_{lm}^{\rm obs}$ and we obtain

$$a_{lm}^{\rm obs} = a_{lm} B_l + \eta_{lm}, \tag{14.38}$$

where there is no summation over $l$ here. The beam effect is a convolution in real space, which in multipole- (or $lm$-) space turns into a simple multiplication if the beam is constant and isotropic. For a Gaussian beam pattern, we have $B_l = \exp(-l^2\theta_{\rm beam}^2/2)$ where $\theta_{\rm beam}$ is related to the full-width half-maximum of the beam. In general, the Fourier transform $B_l$ of the beam is close to 1 on large scales (when $l\theta_{\rm beam} \ll 1$), while it decays to zero on small scales. This corresponds to the fact that the beam washes out anisotropies on scales smaller than $\theta_{\rm beam}$. The noise is simply additive both in the map and in $lm$-space.

In order to turn the data, $a_{lm}^{\rm obs}$, into an estimate $\hat{C}(l)$ of the underlying angular power spectrum to compare to theory, we first ask, what is the probability of getting the data (the $a_{lm}^{\rm obs}$) given the theory (the true $a_{lm}$)? This is analogous to Eq. (14.2). If we assume the noise

has mean zero and a power spectrum given by

$$\left\langle \eta_{lm}\eta^*_{l'm'}\right\rangle = N(l)\delta_{ll'}\delta_{mm'}, \tag{14.39}$$

then this probability is, for a given multipole moment $lm$ (see Exercise 14.5)

$$P(a^{\rm obs}_{lm}|a_{lm}) = \frac{1}{\sqrt{2\pi N(l)}}\exp\left[-\frac{1}{2N(l)}|a^{\rm obs}_{lm} - B_l a_{lm}|^2\right]. \tag{14.40}$$

That is, for fixed true $a_{lm}$, the observed $a^{\rm obs}_{lm}$ follow a multivariate Gaussian distribution with mean $B_l a_{lm}$ (since the noise averages to zero), and variance given by the noise variance $N(l)$. The quantity that we are trying to determine, given a measurement of all $2l+1$ of the $a^{\rm obs}_{lm}$, is $C(l)$, i.e. the underlying true power spectrum of the CMB anisotropies. To derive this, we treat the $a_{lm}$ as random variables, which we have to integrate over. Their distribution $P(a_{lm}|C(l))$ is what we derived in Sect. 9.5.2. So, we write

$$P\left(\{a^{\rm obs}_{lm}\}|C(l)\right) = \prod_{m=-l}^{l}\int da_{lm}\, P(a^{\rm obs}_{lm}|a_{lm})P(a_{lm}|C(l)). \tag{14.41}$$

You can think of the integrand on the right-hand side as $P(a^{\rm obs}_{lm}, a_{lm}|C(l))$, the probability of obtaining the true $a_{lm}$ and observed $a^{\rm obs}_{lm}$ given the underlying $C(l)$. Then, since we have no way of knowing the true $a_{lm}$, we have to marginalize over them.

Since $P(a_{lm}|C(l))$ is a Gaussian with mean zero and variance $C(l)$, we can carry out the integrals, which leads to

$$\mathcal{L} \equiv P\left(\{a^{\rm obs}_{lm}\}|C(l)\right) = \left(2\pi\left[C(l)B_l^2 + N(l)\right]\right)^{-(2l+1)/2}\exp\left\{-\frac{1}{2}\sum_{m=-l}^{l}\frac{\left|a^{\rm obs}_{lm}\right|^2}{C(l)B_l^2 + N(l)}\right\}. \tag{14.42}$$

Armed with this likelihood, we can obtain an estimator for the two-point function, the $C(l)$, and the error on this estimator. The first simply requires us to maximize the likelihood, or more easily, its logarithm, with respect to $C(l)$. The first derivative of the log of the likelihood is

$$\frac{d\ln\mathcal{L}}{dC(l)} = -\frac{(2l+1)B_l^2/2}{C(l)B_l^2 + N(l)} + \frac{1}{2}\sum_{m=-l}^{l}\frac{\left|a^{\rm obs}_{lm}\right|^2 B_l^2}{[C(l)B_l^2 + N(l)]^2}. \tag{14.43}$$

Setting this equal to zero leads to an estimator for $C(l)$:

$$\hat{C}(l) = B_l^{-2}\left(\frac{1}{2l+1}\sum_{m=-l}^{l}\left|a^{\rm obs}_{lm}\right|^2 - N(l)\right). \tag{14.44}$$

We can estimate the error on this estimator the same way we calculated the variance of our estimate of $\sigma_w^2$ in Eq. (14.21). The variance of $\hat{C}(l)$ is

$$\mathrm{Var}\left[\hat{C}(l)\right] = \left\langle \hat{C}(l)^2 \right\rangle - C(l)^2. \tag{14.45}$$

Let us expand the first term on the right and use the fact that, from Eqs. (14.38)–(14.39), $\langle |a_{lm}^{\rm obs}|^2 \rangle = C(l)B_l^2 + N(l)$, so that

$$\left\langle B_l^{-4}\left(\frac{1}{2l+1}\sum_{m=-l}^{l}\left|a_{lm}^{\rm obs}\right|^2 - N(l)\right)^2\right\rangle - C(l)^2 = \left\langle B_l^{-4}\left(\frac{1}{2l+1}\sum_{m=-l}^{l}\left|a_{lm}^{\rm obs}\right|^2\right)^2\right\rangle$$
$$- 2B_l^{-4}N(l)\left(C(l)B_l^2 + N(l)\right) + B_l^{-4}N(l)^2 - C(l)^2.$$

The last line is equal to $-(C(l) + N(l)B_l^{-2})^2$. Given the distribution in Eq. (14.42), the first term on the right is

$$\left\langle B_l^{-4}\left(\frac{1}{2l+1}\sum_{m=-l}^{l}\left|a_{lm}^{\rm obs}\right|^2\right)^2\right\rangle = \frac{2l+3}{2l+1}\left[C(l) + N(l)B_l^{-2}\right]^2 \tag{14.46}$$

so that the error on the estimator for $C(l)$, Eq. (14.44), is

$$\sqrt{\mathrm{Var}\left[\hat{C}(l)\right]} = \sqrt{\frac{2}{2l+1}}\left[C(l) + N(l)B_l^{-2}\right]. \tag{14.47}$$

More precisely, the covariance of the angular power spectrum we should insert in Eq. (14.23) is *diagonal*, as you can verify by calculating $\left\langle \hat{C}(l)\hat{C}(l') \right\rangle$ with $l \neq l'$ and using Eq. (14.39). So we have

$$\mathrm{Cov}_{ll'} = \frac{2}{2l+1}\left[C(l) + N(l)B_l^{-2}\right]^2 \delta_{ll'}. \tag{14.48}$$

The second term in square brackets is the variance due to the noise in the map, amplified by the inverse beam, so that it blows up at large $l$ where the Fourier transform of the beam decays to zero. Even without any noise, the variance does not vanish, since the first term remains. This corresponds to the fundamental uncertainty due to the finite number of $a_{lm}$ on the sky that we can use to obtain an estimate of the variance $C(l)$. Indeed, both terms are downweighted by the number of modes used, $2l+1$. If an experiment does not observe the full $4\pi$ of the sky, but only a fraction $f_{\rm sky}$ of it, then the variance $\mathrm{Cov}_{ll'}$ is approximately increased by a factor $1/f_{\rm sky}$.

These features are quite general for two-point function estimates: the uncertainty is reduced as more modes are measured; there is a noise term, which can be reduced by building more sensitive experiments, but also a *cosmic variance* term due to the finite number of samples. Typically, as in this case, the noise term dominates on small scales (where $B_l^{-2} \gg 1$), while cosmic variance dominates on large scales.

### 14.4.2 Galaxy power spectrum

Much of the previous section carries over to the case of three-dimensional galaxy clustering, but there are some significant differences that are worth calling out. Instead of measuring the angular power spectrum $C(l)$, by (essentially) squaring and averaging the $a_{lm}^{\rm obs}$, we want to measure the 3D galaxy power spectrum by squaring and averaging the Fourier amplitudes $\delta_{\rm g,obs}(\boldsymbol{k})$. So let us first think about how we measure $\delta_{\rm g,obs}(\boldsymbol{k})$. In the following derivation, since we will not encounter any "true" galaxy overdensity, we will let $\delta_{\rm g,obs} \to \delta_{\rm g}$ for clarity.

First, we will assume that the survey consists of a cubic volume with comoving length $L$ on a side. No such surveys exist of course, but this assumption makes the measurement of $\delta_{\rm g}(\boldsymbol{k})$ straightforward. Moreover, this case directly applies to simulated density fields (Sect. 12.3). Imagine covering the three-dimensional survey volume with a cubic grid with $K_{\rm grid}^3$ grid points (the number of grid points determines the maximum value of $|\boldsymbol{k}|$ which we can measure, but the precise number will not be relevant in the following). Then, we construct the density field on the grid $\delta_{\rm g}(\boldsymbol{x}_i) = \delta_{{\rm g},i}$ following Eq. (14.34).

The discrete Fourier transform of the galaxy density field is

$$\delta_{\rm g}(\boldsymbol{k}) = L^{3/2} \sum_i^{K_{\rm grid}^3} \delta_{\rm g}(\boldsymbol{x}_i) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_i}, \quad \text{where} \quad \boldsymbol{k} \in (n_x, n_y, n_z)\, k_F\,, \tag{14.49}$$

and

$$k_F \equiv \frac{2\pi}{L} \tag{14.50}$$

is the wavenumber of the fundamental mode, which precisely covers the box with a single full period. $(n_x, n_y, n_z)$ is a set of whole numbers running from $-K_{\rm grid}/2$ to $K_{\rm grid}/2$. The prefactor of $L^{3/2}$ is chosen for later convenience. The inverse Fourier transform then is

$$\delta_{\rm g}(\boldsymbol{x}) = \frac{1}{K_{\rm grid}^3 L^{3/2}} \sum_{\boldsymbol{k}}^{k_{\rm Ny}} \delta_{\rm g}(\boldsymbol{k}_i) e^{i\boldsymbol{k}_i\cdot\boldsymbol{x}}, \tag{14.51}$$

where the sum over $\boldsymbol{k}$ runs up to the Nyquist frequency of the grid, $k_{\rm Ny} \equiv K_{\rm grid} k_F/2$. The discreteness of the Fourier modes in Eqs. (14.49)–(14.51) is important, because it encodes cosmic variance: the fact that we only have a finite number of Fourier modes available. In the CMB, this discreteness was present from the beginning in the multipole decomposition, since the area of the sky is finite; here, it is due to the finite volume $V = L^3$ of the survey.

Let us then bin the Fourier modes into equally spaced bins $\alpha$ in the magnitude of $\boldsymbol{k}$, so that bin $\alpha$ contains all modes with $k_\alpha - \Delta k/2 \le |\boldsymbol{k}| < k_\alpha + \Delta k/2$. We denote the number of modes in this bin with $m_{k,\alpha}$. Hence, the power spectrum $\hat{P}_{\rm g}(k_\alpha)$ is estimated by averaging over $m_{k,\alpha}$ modes, and this number will play an important role in the error on $\hat{P}_{\rm g}$. So let us count the number of modes. The volume in Fourier space of a spherical shell around $k_\alpha$

is $4\pi k_\alpha^2 \Delta k$, and the number of modes that this volume contains is obtained by dividing by the volume of the fundamental cell in Fourier space, which is given by $k_F^3$:

$$m_{k,\alpha} = \frac{1}{2}\frac{4\pi k_\alpha^2 \Delta k}{k_F^3}. \tag{14.52}$$

The factor of $1/2$ in front is due to the fact that the density field is real, so that only one half of the Fourier modes are actually independent, with the other half being fixed by the reality condition $\delta_{\rm g}(-\boldsymbol{k}) = \delta_{\rm g}^*(\boldsymbol{k})$. Therefore,

$$m_{k,\alpha} = \frac{1}{4\pi^2} V k_\alpha^2 \Delta k. \tag{14.53}$$

The estimator for the galaxy power spectrum then is directly analogous to Eq. (14.44). Moreover, we do not have to include a beam or point-spread function in the case of galaxy redshift surveys, since the angular resolution of telescopes is extremely high. While the grid itself has finite resolution, this is just a numerical tool and we can always increase this resolution if necessary. So, translating Eq. (14.44) to the 3D case and setting $B_l \to 1$, we have

$$\hat{P}_{\rm g}(k_\alpha) = \frac{1}{m_{k,\alpha}} \sum_{\boldsymbol{k}}^{||\boldsymbol{k}|-k_\alpha|<\Delta k/2} |\delta_{\rm g}(\boldsymbol{k})|^2 - P_N, \tag{14.54}$$

where $P_N$ is the noise. For Poisson noise, usually assumed for forecasts, $P_N = \bar{n}_{\rm g}^{-1}$; in general, the noise needs to be determined from the data. The error on $\hat{P}_{\rm g}$ likewise follows analogously to the CMB case, Eq. (14.47). Its derivation is instructive though, so let us briefly go through it here. The covariance is defined as

$$\begin{aligned}
{\rm Cov}_{\alpha\beta} &\equiv \left\langle \hat{P}_{\rm g}(k_\alpha)\hat{P}_{\rm g}(k_\beta)\right\rangle - \left\langle \hat{P}_{\rm g}(k_\alpha)\right\rangle \left\langle \hat{P}_{\rm g}(k_\beta)\right\rangle \\
&= \frac{1}{m_{k,\alpha}} \sum_{\boldsymbol{k}}^{||\boldsymbol{k}|-k_\alpha|<\Delta k/2} \frac{1}{m_{k,\beta}} \sum_{\boldsymbol{k}'}^{||\boldsymbol{k}'|-k_\beta|<\Delta k/2} \left[\left\langle |\delta_{\rm g}(\boldsymbol{k})|^2 |\delta_{\rm g}(\boldsymbol{k}')|^2\right\rangle - \left\langle |\delta_{\rm g}(\boldsymbol{k})|^2\right\rangle \left\langle |\delta_{\rm g}(\boldsymbol{k}')|^2\right\rangle\right],
\end{aligned} \tag{14.55}$$

where in the second line we have inserted Eq. (14.54) and dropped $P_N$; since $P_N$ is just a constant, it can be pulled out of the expectation values and disappears from Eq. (14.55) (as you should check and convince yourself of). Let us look at the first expectation value in Eq. (14.55), which involves four instances of $\delta_{\rm g}$. We can expand it using Wick's theorem (see Box 12.1):

$$\begin{aligned}
\left\langle |\delta_{\rm g}(\boldsymbol{k})|^2 |\delta_{\rm g}(\boldsymbol{k}')|^2\right\rangle &= \left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(-\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\delta_{\rm g}(-\boldsymbol{k}')\right\rangle \\
&= \left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(-\boldsymbol{k})\right\rangle \left\langle \delta_{\rm g}(\boldsymbol{k}')\delta_{\rm g}(-\boldsymbol{k}')\right\rangle \\
&\quad + \left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\right\rangle \left\langle \delta_{\rm g}(-\boldsymbol{k})\delta_{\rm g}(-\boldsymbol{k}')\right\rangle \\
&\quad + \left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(-\boldsymbol{k}')\right\rangle \left\langle \delta_{\rm g}(-\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\right\rangle \\
&\quad + \left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(-\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\delta_{\rm g}(-\boldsymbol{k}')\right\rangle_{\rm conn}.
\end{aligned} \tag{14.56}$$

The four terms here correspond to the three terms obtained from Wick's theorem, augmented with the last term labeled with a subscript $_{\rm conn}$: this "connected" term is only present if the field $\delta_{\rm g}$ is not Gaussian. Let us ignore it for the time being; we will discuss its significance below.

The first term in Eq. (14.56) simply cancels the second term in the covariance in Eq. (14.55), so we only need to consider the second and third term in Eq. (14.56). For our definition of the discrete Fourier-space field (Eq. (14.49)), the power spectrum is given by

$$\left\langle \delta_{\rm g}(\boldsymbol{k})\delta_{\rm g}(\boldsymbol{k}')\right\rangle = \delta_{\boldsymbol{k},-\boldsymbol{k}'}\left[P_{\rm g}(k)+P_N\right], \tag{14.57}$$

where the Kronecker symbol $\delta_{\boldsymbol{k},-\boldsymbol{k}'}$ is unity if $\boldsymbol{k}$ and $-\boldsymbol{k}'$ are equal, and zero otherwise; recall that the components of $\boldsymbol{k}$ and $\boldsymbol{k}'$ are integer multiples of $k_F$. Notice that the simple forms of Eq. (14.54) and Eq. (14.57), obtained simply from the continuous version Eq. (C.22) by replacing

$$(2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}') \to \delta_{\boldsymbol{k},-\boldsymbol{k}'}, \tag{14.58}$$

hold thanks to the $L^{3/2}$ factor we included in the definition of $\delta_{\rm g}(\boldsymbol{k})$. Comparing with Eq. (14.55), we see that the second and third term in Eq. (14.56) can only contribute if the wavenumber bins $\alpha$ and $\beta$ overlap; so, considering the standard case of non-overlapping bins, this means that the covariance is only nonzero if $\alpha=\beta$. Moreover, considering a single term in the sum over $\boldsymbol{k}$ in Eq. (14.55) in the $\alpha=\beta$ case, there are only two values, $\boldsymbol{k}'=\boldsymbol{k}$ and $\boldsymbol{k}'=-\boldsymbol{k}$, that contribute to the sum over $\boldsymbol{k}'$ (recall that the sum is over all wavenumber directions, with the magnitude of $\boldsymbol{k}$, $\boldsymbol{k}'$ constrained to be within the chosen bin). So, if $\alpha=\beta$, the sums in Eq. (14.55) yield $2/m_{k,\alpha}$ multiplied by the power spectrum (plus noise) squared. The covariance of the galaxy power spectrum thus is given by

$$\mathrm{Cov}_{\alpha\beta} = \frac{2}{m_{k,\alpha}}\left[P_{\rm g}(k_\alpha)+P_N\right]^2\delta_{\alpha\beta}. \tag{14.59}$$

This is very analogous to our CMB result, Eq. (14.48), the only difference being the number of modes ($m_{k,\alpha}$ vs. $2l+1$), and the absence of the beam. An easy follow-up result is the error on $P_{\rm g}(k_\alpha)$, given by the square root of the diagonal covariance element:

$$\sqrt{\mathrm{Var}\left[\hat{P}_{\rm g}(k_\alpha)\right]} = \sqrt{\frac{2}{m_{k,\alpha}}}\left[P_{\rm g}(k_\alpha)+P_N\right]. \tag{14.60}$$

Even for an infinitely dense galaxy sample, so that $P_N\to 0$, a minimum error remains, which is the sample variance due to the finite number of modes available in the survey volume $V$ (Eq. (14.53)). For simplicity we have here considered a wavevector bin that averages over all directions of $\boldsymbol{k}$. This means that we are actually estimating the monopole part of the galaxy power spectrum. One can also measure the higher multipoles $l>0$ of the anisotropic galaxy power spectrum Eq. (11.23) by choosing a weighting based on $\mathcal{P}_l(\mu)$ in Eq. (14.54), where $\mu=\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{n}}$. Alternatively, one can divide each wavenumber bin $\alpha$ into multiple bins of $\mu$.

Eq. (14.60) together with Eq. (14.53) shows that the measurement precision of the galaxy power spectrum improves rapidly with increasing $k_\alpha$ (in the case of equally spaced linear bins $\Delta k$, $\sqrt{2/m_{k,\alpha}} \propto 1/k_\alpha$), so long as $P_{\rm g}(k_\alpha)$ is not much smaller than the noise $P_N$. There are two challenges involved in exploiting this great amount of information apparently available on small scales. First, as we learned in Ch. 12, coming up with a theoretical prediction for the galaxy power spectrum becomes increasingly difficult on small scales, due to the importance of the nonlinearities in matter and galaxy bias. Second, once the galaxy density field becomes moderately nonlinear, the Gaussian assumption that our derivation was built on becomes inaccurate. That is, Eqs. (14.59)–(14.60) receive an additional contribution from the last, connected term in Eq. (14.56) which is induced by nonlinear evolution (it can be calculated, for example, using perturbation theory along the lines of Sect. 12.2 and Sect. 12.6). Unlike the Gaussian contribution, the connected covariance term couples different $k$ bins. Due to these challenges, the analysis of the galaxy power spectrum is usually restricted to scales $k_\alpha \leq k_{\rm max}$, where $k_{\rm max}$ is typically chosen to be on scales where perturbation theory is valid ($k_{\rm max} \lesssim 0.2\,h\,{\rm Mpc}^{-1}$, depending on redshift).

So far, we assumed a cubic survey volume (and one with periodic boundary conditions, which are built into the discrete Fourier transform of Eq. (14.49)). In order to move to realistic survey geometries, we go back to the discussion around Eq. (14.34), where we described how the actual region of space covered by a given survey is embedded in a larger cubic volume. The cubic volume considered here thus corresponds to the grid in which the actual survey is embedded. Eq. (14.34) then only applies to those cells of the grid that are actually in the observed volume. So, we can write

$$\delta_{\rm g}^{\rm obs}(\boldsymbol{x}_i) = \mathcal{W}(\boldsymbol{x}_i)\delta_{\rm g}(\boldsymbol{x}_i), \tag{14.61}$$

where $\delta_{\rm g}$ is the galaxy density that would be measured if all pixels were actually part of the survey, and $\mathcal{W}(\boldsymbol{x}_i)$ is the window function. In the simplest case, the window function only attains two values: 1, if $\boldsymbol{x}_i$ is in the surveyed volume, and 0 otherwise. Very roughly, the region where $\mathcal{W}$ is unity has the form of a truncated cone, limited in the transverse directions by the survey footprint on the sky, and in the line-of-sight direction by the minimum and maximum redshift of the survey. In reality, the mask usually has some small holes as well, due to bright foreground stars which need to be masked, as well as other observational effects.

In Fourier space, Eq. (14.61) becomes

$$\delta_{\rm g}^{\rm obs}(\boldsymbol{k}_i) = \sum_{\boldsymbol{k}_j} \mathcal{W}(\boldsymbol{k}_j)\delta_{\rm g}(\boldsymbol{k}_i - \boldsymbol{k}_j). \tag{14.62}$$

The likelihood for this observed, windowed density field is still Gaussian (or as close to Gaussian as that of $\delta_{\rm g}$ itself), but the covariance no longer takes the simple form of Eq. (14.59): the window function couples different Fourier modes, so that the covariance attains a complicated, non-diagonal form. Apart from this complication, however, the

same principle to construct a maximum-likelihood estimator for the galaxy power spectrum applies as before.

## 14.5 The Fisher matrix

As we have seen in the previous sections, obtaining cosmological constraints from data is a multi-step process that depends in great detail on the properties of the data set considered. However, there is a much simpler question that can be answered even before obtaining the data: how well do we expect a given experiment to determine cosmological parameters? This exercise is known as *forecasting*. All we need for this purpose is the curvature matrix of the likelihood, the *Fisher matrix*, which quantifies the amount of information that a given experiment can provide about a set of parameters. Moreover, we can readily compute it using the results of the previous sections, without having to work with actual data sets. The Fisher matrix has thus become a useful tool for theorists as well, allowing them to determine whether a new signal they predict could in fact be detected experimentally.

Let us consider a CMB experiment as an example. Start with the following:

- The set of cosmological parameters $\{\lambda_\alpha\}$ for which we want to forecast errors, and their fiducial values $\{\bar{\lambda}_\alpha\}$ which are assumed to describe the true universe.
- A theory prediction, $C^{\rm theory}(l|\lambda_\alpha)$, as a function of cosmological parameters $\{\lambda_\alpha\}$.
- The expected uncertainty, $\mathrm{Var}\left[\hat{C}(l)\right]$, on the $C(l)$ expected from a given experiment.

The observed $\hat{C}(l)$ in this experiment will be close to, within errors, the true $C^{\rm theory}(l)$; indeed, if we form

$$\chi^2(\{\lambda_\alpha\}) = \sum_l \frac{[\hat{C}(l) - C^{\rm theory}(l, \{\lambda_\alpha\})]^2}{\mathrm{Var}\left[\hat{C}(l)\right]}, \tag{14.63}$$

where we have assumed a diagonal covariance for simplicity, then we expect $\chi^2$ to reach a minimum at the point in parameter space where $\{\lambda_\alpha\} = \{\bar{\lambda}_\alpha\}$, the values of the parameters that we assume to describe the true universe. Of course, we do not know what those values are for the real universe, but even without that information, we can ask how quickly $\chi^2(\{\lambda_\alpha\})$ changes as a given parameter $\lambda_1$ moves away from $\bar{\lambda}_1$. If it increases rapidly, then the error on the parameter will be very small; if $\chi^2$ changes little, then there will be a large error on $\lambda_1$.

To quantify this, we can expand $\chi^2$ about its minimum at $\bar{\lambda}_\alpha$. Let us first do this in the case of one parameter; the generalization to many parameters will be straightforward. In the one-parameter case,

$$\chi^2(\lambda) = \chi^2(\bar{\lambda}) + \mathcal{F}(\lambda - \bar{\lambda})^2. \tag{14.64}$$

The linear term in Eq. (14.64) vanishes since $\chi^2$ is at a minimum at $\bar{\lambda}$. The coefficient of the quadratic term is

$$\mathcal{F} \equiv \frac{1}{2}\frac{\partial^2\chi^2}{\partial\lambda^2}. \tag{14.65}$$

The curvature here, $\mathcal{F}$, measures how rapidly $\chi^2$ changes away from its minimum. If the curvature is small, then the likelihood changes slowly and the data are not very constraining: the resulting uncertainties on the parameter will be large. Conversely, large curvature translates into small uncertainties. Since we are assuming that the likelihood for $\hat{C}(l)$ is Gaussian, so that $\ln\mathcal{L} = -\chi^2/2$, $\mathcal{F}$ is the generalization of our error estimate for $\bar{w}$ in Eq. (14.12). Therefore, the 1-$\sigma$ error on $\lambda$ is indeed simply $1/\sqrt{\mathcal{F}}$.

The second derivative of $\chi^2$ contains two terms:

$$\mathcal{F} = \sum_l \frac{1}{\mathrm{Var}\left[\hat{C}(l)\right]}\left[\left(\frac{\partial C^{\mathrm{theory}}(l,\lambda)}{\partial\lambda}\right)^2 + (C^{\mathrm{theory}}(l,\lambda) - \hat{C}(l))\frac{\partial^2 C^{\mathrm{theory}}(l,\lambda)}{\partial\lambda^2}\right], \tag{14.66}$$

where, as throughout this chapter, we have assumed that the covariance of $\hat{C}(l)$ is independent of $\lambda$. We are interested in computing the expected uncertainty on $\lambda$, i.e. what a given experiment would yield if repeated many times. So, we should take the *expectation value* of $\mathcal{F}$, which actually leads to further simplifications: under that expectation value, $\hat{C}(l) - C^{\mathrm{theory}}(l,\bar{\lambda})$, which is nonzero in any given realization due to noise and cosmic variance, vanishes by definition at the true parameter value $\bar{\lambda}$ (note that the only random variable here is $\hat{C}(l)$). Thus, only the first term remains under the expectation value:

$$F \equiv \langle\mathcal{F}\rangle = \sum_l \frac{1}{\mathrm{Var}\left[\hat{C}(l)\right]}\left[\frac{\partial C^{\mathrm{theory}}(l,\bar{\lambda})}{\partial\bar{\lambda}}\right]^2 \tag{14.67}$$

Notice that this now contains only "theory" quantities: the data vector does not appear. This ensemble average of the curvature of the likelihood at the maximum (i.e. at the true or fiducial parameter values) is called the *Fisher information* $F$. The generalization of this quantity to many parameters is called the Fisher information matrix, or *Fisher matrix* for short:

$$F_{\alpha\beta} = \sum_l \frac{1}{\mathrm{Var}\left[\hat{C}(l)\right]}\frac{\partial C^{\mathrm{theory}}(l,\{\bar{\lambda}_\gamma\})}{\partial\bar{\lambda}_\alpha}\frac{\partial C^{\mathrm{theory}}(l,\{\bar{\lambda}_\gamma\})}{\partial\bar{\lambda}_\beta}. \tag{14.68}$$

In order to predict how accurately parameters will be known, then, we simply need to know the experiment's specifications (to determine $\mathrm{Var}[\hat{C}(l)]$) and the derivatives of the $C^{\mathrm{theory}}(l,\{\bar{\lambda}_\alpha\})$ around their assumed true values (which are typically evaluated numerically using a finite-difference estimate). Eq. (14.68) assumes a Gaussian likelihood and a diagonal, $\{\lambda_\alpha\}$-independent covariance. The generalization to *any* likelihood shape is given by

$$F_{\alpha\beta} \equiv -\left\langle\frac{\partial^2\ln\mathcal{L}}{\partial\lambda_\alpha\partial\lambda_\beta}\right\rangle\Bigg|_{\{\lambda_\gamma\}=\{\bar{\lambda}_\gamma\}}. \tag{14.69}$$

For a Gaussian distribution and a single parameter $\lambda$, the forecasted 1-$\sigma$ uncertainty on $\lambda$ is $1/\sqrt{F}$. How about if more than one parameter is allowed to vary? Fig. 14.4 illustrates

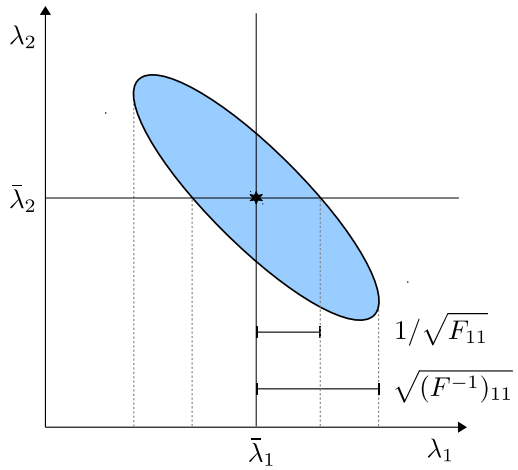



**FIGURE 14.4** Error ellipse in a 2D parameter space. The maximum of the likelihood $(\bar{\lambda}_1, \bar{\lambda}_2)$ is indicated by a star. Two possible errors on the parameter $\lambda_1$ are shown: if $\lambda_2$ is perfectly known (infinitely narrow prior), then the error on $\lambda_1$ corresponds to the width of the ellipse at $\lambda_2 = \bar{\lambda}_2$, equal to $1/\sqrt{F_{11}}$. If no prior information is available about $\lambda_2$, then the proper error to quote is the width of the distribution after marginalizing over $\lambda_2$, which yields the larger error bar given by $\sqrt{(F^{-1})_{11}}$.

the situation in a two-dimensional setting. If the parameter $\lambda_2$ is assumed known, then the error on $\lambda_1$ is $1/\sqrt{F_{11}}$. However, if $\lambda_2$ is allowed to vary, the proper error on $\lambda_1$ is obtained after integrating over all possible values of $\lambda_2$, which yields $\sqrt{(F^{-1})_{11}}$. It is instructive to prove this explicitly. First, using the fact that the Fisher matrix describes the curvature of the likelihood (in terms of the $\lambda_\alpha$) around the maximum, we can write the joint posterior for the two parameters as

$$P(\lambda_1, \lambda_2) \propto \exp\left\{-\frac{1}{2}\lambda_\alpha F_{\alpha\beta}\lambda_\beta\right\} \tag{14.70}$$

where we have assumed that the true values are $(\lambda_1, \lambda_2) = (0, 0)$ for simplicity. Allowing $\lambda_2$ to vary is equivalent to integrating (marginalizing) this probability distribution over all possible values of $\lambda_2$, as in Eq. (14.22). Then

$$\begin{aligned} P(\lambda_1) &= \int d\lambda_2\, P(\lambda_1, \lambda_2) \\ &\propto \exp\left\{-\frac{\lambda_1^2}{2}\left(\frac{F_{11}F_{22} - F_{12}F_{21}}{F_{22}}\right)\right\} \end{aligned} \tag{14.71}$$

where the second line comes from carrying out the $\lambda_2$ integration explicitly. The term in parentheses is $[F_{11}F_{22} - F_{12}F_{21}]/F_{22} = 1/(F^{-1})_{11}$. So the $1\sigma$ error is indeed given by $\sqrt{(F^{-1})_{11}}$.

## 14.6 Sampling the likelihood function

We have gone through all the ingredients needed for the likelihood in this Chapter: the theory predictions, estimates of the two-point functions from the data, and their covariance. Now we want to constrain cosmological parameters using this likelihood. Finding the maximum of the likelihood analytically, like we did for the estimators of the map and two-point functions, is not generally possible for several reasons. First, as discussed below Eq. (14.23), the likelihood is in general not close to Gaussian in the cosmological parameters. In addition, we usually have to integrate over several nuisance parameters which can rarely be done analytically. Instead, we need to proceed numerically.

In principle, then, we can go ahead in a brute-force approach and compute this likelihood function at many points in parameter space, find its maximum (this constitutes the set of best-fit parameters), and the contour delineating the region in which, say, 95% of the volume lies around this maximum. This contour would then be the 95% confidence region of the parameters. This brute-force approach, however, is entirely impractical for the multi-dimensional parameter sets needed for modern experiments. The typical number of parameters required to describe cosmological data is of order ten to one hundred. Say that mapping the likelihood in a one-dimensional parameter space requires 20 likelihood evaluations; in 2D, this will become $20^2$; for three parameters, $20^3$, etc. When there are 20 free parameters, the number of likelihood evaluations required would be $20^{20}$. Even if the evaluation of the likelihood were to only take a few seconds (in reality, the likelihood function is often very costly to evaluate), it is easy to compute that this brute-force approach is practically unfeasible.

Thus, we need new techniques for evaluating the likelihood function and finding its maximum and its width. One of the boons of the tools we will discuss here is that they are applicable to non-Gaussian likelihoods as well, which is usually the case in real-world applications.

Fundamentally, the issue we are facing in complex likelihood analyses is that we have to find probable regions in a generally high-dimensional parameter space $\{\lambda_\alpha\}$. Moreover, after having found the maximum, we have to perform integrals over the likelihood to obtain the proper marginalized error bars on individual parameters (Sect. 14.5). As argued above, we need to be smart when attempting to do this in practice.

Suppose we had an algorithm that, given any posterior—the product of the likelihood function and priors—returns us points ("samples") in parameter space $\{\lambda_\alpha^i\}_{i=1}^{m_{\rm sample}}$ that are statistically independent from each other and whose distribution follows the posterior. Fig. 14.5 shows a one-dimensional example. Then our problem would be solved: the desired best-fit parameter $\lambda_\alpha$ is given by the mean of the samples,

$$\bar{\lambda}_\alpha = \frac{1}{m_{\rm sample}} \sum_{i=1}^{m_{\rm sample}} \lambda_\alpha^i, \tag{14.72}$$

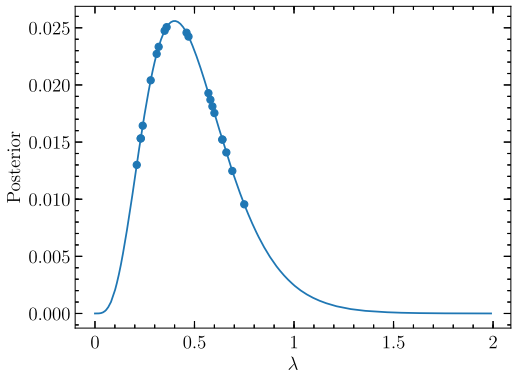



**FIGURE 14.5** The posterior distribution of a parameter $\lambda$ (solid curve). The dots are points sampled from that distribution, so that values of $\lambda$ where the posterior is large are chosen more often. With a sufficient number of samples, the mean and the variance of the sampled points then are close to the true values of the distribution. In this case, the true mean and standard deviation of the posterior are 0.5 and 0.22, while the mean of the 20 points sampled from the distribution is equal to 0.46, with a standard deviation of 0.18. If more points were sampled, the mean and standard deviation would continue to approach the true values.

where the sum runs over all $m_{\text{sample}}$ samples, while the marginalized error can be estimated as the sample variance of $\lambda_\alpha^i$,

$$\text{Var}\,[\lambda_\alpha] = \frac{1}{m_{\text{sample}} - 1} \sum_{i=1}^{m_{\text{sample}}} \left(\lambda_\alpha^i - \bar{\lambda}_\alpha\right)^2 . \tag{14.73}$$

In the limit that the number of samples is very large, $\bar{\lambda}_\alpha$ and $\text{Var}\,[\lambda_\alpha]$ converge to central values that are independent of the number of samples. In fact, one could do much more: given a sufficiently large $m_{\text{sample}}$, a normalized histogram of $\lambda_\alpha^i$ yields the marginalized probability distribution of $\lambda_\alpha$. Since this is a one-dimensional function, it is then easy to find its global maximum and derive proper confidence intervals. Similarly, we could obtain the joint probability distribution of two parameters $(\lambda_\alpha, \lambda_\beta)$ to see whether they are degenerate, for example. Clearly then, if numerically efficient, this approach would be a solution to our problems.

Fortunately, such algorithms exist; the most popular is known as *Markov Chain Monte Carlo (MCMC)*. The "Monte Carlo" refers to the fact that we will be throwing dice (i.e. using a random number generator) in the process. "Markov Chain" means that, to generate sample $i + 1$, the algorithm only uses the previous sample point $\lambda^i$ as input (in addition to random numbers). This is a significant restriction: it means that the algorithm has no memory, i.e. it does not care which previous samples $\lambda^1, \lambda^2, \ldots$, also known as *chain*, took us to $\lambda^i$. Here and in the following, we will let $\lambda$ stand for the parameter vector $\{\lambda_\alpha\}$. Let us begin by deriving what the condition is for this algorithm to work, i.e. to actually yield sam-

ples that follow the desired posterior $P(\lambda)$.[3] Our main goal then is to identify a candidate algorithm that inputs the current value of the parameter set, $\lambda$, and outputs a subsequent set, call it $\lambda'$.

Given its Markov nature, the algorithm is completely described by the conditional probability $K(\lambda'|\lambda)$ that takes us from a sample $\lambda$ to the next one, $\lambda'$. The fundamental requirement on $K$, in order for the MCMC sampler to sample from the right posterior, is *detailed balance*:

$$P(\lambda)K(\lambda'|\lambda) = P(\lambda')K(\lambda|\lambda'). \tag{14.74}$$

If we start with a distribution of $\lambda$ that follows $P(\lambda)$, then an algorithm that obeys Eq. (14.74) preserves that distribution. This is analogous to detailed balance in the case of the collision terms in the Boltzmann equation, which preserve the equilibrium distribution: writing Eq. (14.74) in the form $P(\lambda)K(\lambda'|\lambda) - P(\lambda')K(\lambda|\lambda') = 0$ shows that it corresponds to the statement that the rates for forward ($\lambda \to \lambda'$) and reverse reactions ($\lambda' \to \lambda$) are the same (see for example the discussion in Sect. 5.1). Hence, if we want to sample from a posterior $P(\lambda)$, we need to identify a scattering process whose equilibrium distribution is $P(\lambda)$, and then simulate that scattering.

Before thinking about how we can identify such an algorithm, let us first understand why the detailed balance requirement Eq. (14.74) is in fact what we need. It is clear that if we start from a distribution of initial samples that follow the desired posterior, then Eq. (14.74) ensures that we continue to sample from the correct distribution. In practice we will start from some initial guess for $\lambda$ which is likely to be very far from the maximum of the posterior. The reason that Eq. (14.74) still ensures the right result also follows from the Boltzmann analogy: if we start with a photon distribution that is very far from the Bose–Einstein distribution and bring it into contact with electrons at some temperature, it will be driven to the equilibrium distribution with the same temperature after many scatterings. Similarly, even when started with initial samples that are very far from the true posterior distribution, a sampling algorithm that satisfies detailed balance ensures that their distribution will approach $P(\lambda)$ after sufficiently many "scatterings," i.e. random samples. So, the approach to the true distribution does not happen immediately, but during a "burn-in" period.

One choice for $K(\lambda'|\lambda)$ is the Metropolis-Hastings algorithm invented in the 1950s by Metropolis and later generalized by Hastings. Focusing on a single parameter for simplicity, we select a possible next sample $\lambda'$ by drawing from a distribution centered around $\lambda$ and symmetric in its two arguments: $g(\lambda', \lambda)$. The simplest example is a Gaussian, $g(\lambda', \lambda) \propto \exp[-(\lambda - \lambda')^2/2\sigma^2]$. Then, this new sample is "accepted" with a probability

$$p_{\rm acc}(\lambda', \lambda) = \min\left\{\frac{P(\lambda')}{P(\lambda)},\ 1\right\}. \tag{14.75}$$

[3] Really, this is $P(\lambda|\{d_i\})$, but since the data are fixed and do not appear explicitly in this section, we drop them from the argument for clarity.

That is, we evaluate the posterior for $\lambda'$ and compute the ratio $\alpha \equiv P(\lambda')/P(\lambda)$. If $\alpha$ is larger than one, then $\lambda'$ is our next step in the chain; if it is less than one, we uniformly draw a random number between zero and one and accept $\lambda'$ into the chain only if that random number is less than $\alpha$. So, if the proposed $\lambda'$ is much less likely than the previous value $\lambda$, it is unlikely (but not impossible) for it to be accepted. On the other hand, if the random number is greater than $\alpha$, then $\lambda'$ is discarded and we instead insert the previous value $\lambda$ into the chain. We then repeat the cycle, drawing a new proposal based on the last element of the chain.

Let us show that this recipe satisfies detailed balance. This requires that

$$\begin{aligned} \frac{P(\lambda')}{P(\lambda)} &= \frac{K(\lambda'|\lambda)}{K(\lambda|\lambda')} \\ &= \frac{p_{\rm acc}(\lambda',\lambda)}{p_{\rm acc}(\lambda,\lambda')}, \end{aligned} \tag{14.76}$$

where the second line follows since $g(\lambda',\lambda)$ is symmetric. If $P(\lambda') < P(\lambda)$, then the denominator is equal to one, and the numerator is indeed equal to $P(\lambda')/P(\lambda)$. Similarly, if $P(\lambda') > P(\lambda)$, the numerator is equal to one, and the denominator ensures equality, so detailed balanced holds for this algorithm. The algorithm is also simply generalized to multiple parameters, by performing successive steps in $\lambda_1, \lambda_2, \ldots$, either with the same function $g(\lambda_i', \lambda_i)$ for each parameter, or with different functions $g_i$.

Related to this is one advantage of the Metropolis-Hastings algorithm: the function $g$ can be tuned. For example, for the Gaussian choice above, $\sigma$ is a free parameter. If it is chosen to be very small, the sampler will take a long time to map out the likelihood function, or might even get stuck near a local maximum. If $\sigma$ is too large, the acceptance rate will be low, because most draws of $\lambda'$ will end up in low probability regions of parameter space. Hence, during the burn-in phase, the step-size parameter is often adjusted dynamically. A downside of this algorithm is that, due to the rejection sampling, several evaluations of the posterior might be required to generate a new sample. Moreover, successive samples are not truly statistically independent, so care needs to be taken when evaluating the sample mean and variance in Eqs. (14.72)–(14.73).

Metropolis-Hastings was one of the first MCMC algorithms, but since then many others have been developed and are in wide use in cosmology (and of course in the much broader field of data science). Some of the basic issues mentioned above apply to all samplers:

- Step-size optimization to map the likelihood with the fewest number of evaluations
- Estimating the burn-in period
- Measuring when the chain has converged
- Understanding the correlation between adjacent samples.

## 14.7 Summary

This is the second edition of a book that was first published in 2003. Given all the advances in cosmology, it did not need to surprise us that this chapter was one of the ones

that needed the most revision. The field of data analysis has changed rapidly over the past decades, and indeed it is with a bit of trepidation that we present the steps in this chapter, as there are already a number of tools lurking on the horizon—for example, likelihood-free inference, or forward-modeling of the entire map as an alternative to the power spectrum, perhaps coupled with machine learning—that may replace the likelihood analysis outlined here.

Nonetheless, the Gaussian statistics we focused on in this chapter still form the bedrock on which all statistical inference approaches are built. Moreover, even future, more advanced techniques are very likely to rely on one or several of the steps shown in Fig. 14.1. These include two *compression* steps: one that takes the raw data and produces a map and a second that estimates the two-point functions from the map. The theory part, computing the two-point functions in a given model, encompasses essentially all the previous chapters of this book, and will remain relevant to future analyses. There is an additional theory step, obtaining a covariance matrix for the two-point function, that we have described here in the simplest of cases. The diagonal covariance matrices for the $C(l)$ and the galaxy power spectrum we described here are often used for forecasts, but have been inadequate for the analysis of real data for some time now. Covariances are in fact one of the major challenges in the analysis chain of current cosmological data sets. Finally, armed with those three pieces, evaluating the likelihood and therefore obtaining parameter constraints turns into a numerical statistics problem, one for which the most widely-used solution currently is the Markov Chain Monte Carlo.

Before concluding this chapter and this book, we should mention another major challenge: systematic effects. Common examples in the area of galaxy clustering are contamination of the sample by stars in the Milky Way, mis-estimated redshifts, and imperfect theoretical models, for example in the bias relation between the galaxy density and the matter density. In the case of the CMB, galactic foregrounds and instrumental effects such as the beam are prominent examples. One of the most promising ways to include many of these systematics is by forward-modeling. For example, template maps for various sources of galactic foreground emission can be constructed by measuring the sky brightness at different frequencies. Then, one allows for the CMB map to be contaminated by these templates, parametrized by free amplitude coefficients, which are then constrained by the data simultaneously with the cosmological parameters, similar to how we deal with galaxy bias. This, however, only works with known systematics. An array of null tests, estimators constructed to yield null results in the absence of systematics, can be used to search for any other, unknown source of systematics. All of these techniques are somewhat specific to the observable considered (CMB, galaxy clustering, lensing, and so on), and indeed worthy of their own textbook, so we have refrained from going into further detail here.

## Exercises

**14.1** Alice has a cough that persists for a week. A primary symptom of lung cancer is a cough; quantify that by writing $P(C|LC) = 1$. Roughly, 1 in 2000 people in the

United States are diagnosed with lung cancer every year. Use Bayes' theorem to estimate the chances that Alice has developed lung cancer, $P(LC|C)$. Assume that, in a given year, 1 in five people get coughs that last for a week. This is a dramatic illustration of the importance of priors for everyday life.

**14.2** Generalize the likelihood Eq. (14.2) to different errors on each measurement, $\sigma_w \to \sigma_{w,i}$. Derive the value of $w$ that maximizes this likelihood, and show that it corresponds to inverse-variance weighting.

**14.3** Show that the covariance matrix of a map,

$$\left\langle (\hat{s}_i - s_i)(\hat{s}_j - s_j) \right\rangle, \tag{14.77}$$

where $\hat{s}_i$ is given by the estimator in Eq. (14.30), is $C_N$, as given in Eq. (14.31).

**14.4** Derive the effect of a general, position-dependent and anisotropic beam on the $a_{lm}^{\rm obs}$ on the full sky. Begin by writing the observed temperature as in Eq. (14.36). Neglect the noise which is irrelevant for this calculation. Now decompose all quantities into multipole moments, such that $B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}')$ is replaced by $B_{lm,l'm'}$. Show that the observed temperature monopoles are given by

$$a_{lm}^{\rm obs} = \sum_{l'm'} B_{lm,l'm'} a_{l'm'}. \tag{14.78}$$

Now specialize to a constant and isotropic beam $B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') = B(\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{n}}')$ as considered in Sect. 14.4.1. Show that in this case Eq. (14.38) holds.

**14.5** Use the fact that the noise $\eta_{lm}$ is a Gaussian random variable with mean zero and covariance given by Eq. (14.39) to derive Eq. (14.40). Carry out the integral in Eq. (14.41) to obtain the likelihood in Eq. (14.42).

**14.6** Consider an all-sky CMB experiment with spatial pixels of area $\Delta\Omega$. Assume that the experiment measures the temperature in each pixel with Gaussian noise $\sigma_\eta$. The noise amplitude is often given in units of $\mu$K/arcmin$^2$, so that the noise $\sigma_\eta$ on the temperature perturbation $\Theta$ is obtained by multiplying this number by the pixel area $\Delta\Omega$ and dividing by $T_0 = 2.726$ K. The noise is thus assumed to be uniform (the same everywhere on the sky) and uncorrelated (from one pixel to the next). Determine the noise covariance matrix for $a_{lm}^{\rm obs}$. If the pixel area is cut in half (for the same experiment), each pixel will get less observing time by a factor of 2. The noise will then go up for each pixel by a factor of $\sqrt{2}$. Show that these two changes (smaller pixels; more noise per pixel) leave the noise covariance unchanged.

**14.7** Derive the noise contribution to the lensing angular power spectrum $C_{EE}(l)$ (Sect. 13.5). This is analogous to the CMB noise derived in Exercise 14.6, except that the noise is due to the intrinsic ellipticities of galaxies. Start from Eq. (13.40), and treat $\epsilon_i^{\rm true}$ $(i = 1, 2)$ as Gaussian noise with $\langle \epsilon_i^{\rm true} \rangle = 0$ and $\langle (\epsilon_i^{\rm true})^2 \rangle = \sigma_\epsilon^2$. Then, assume that a pixel with area $\Delta\Omega$ contains $\bar{n}_{\rm g}\Delta\Omega$ galaxies, where $\bar{n}_{\rm g}$ is the angular number density of source galaxies (number of galaxies per solid angle).

**14.8** In Sect. 14.4.2, we assumed that the noise contribution to the galaxy covariance is diagonal with amplitude $P_N$. Show that this is the case, and what value $P_N$ attains, assuming that noise in galaxy counts is a Poisson process:

**(a)** Divide the survey region into small sub-volumes. Assume that the number of galaxies in a given sub-volume is drawn from a Poisson distribution with mean $\bar{m}_{\rm g}$ (assume $\bar{m}_{\rm g}$ is the same in all sub-volumes for simplicity),

$$P(m) = \frac{(\bar{m}_{\rm g})^m e^{-\bar{m}_{\rm g}}}{m!}. \tag{14.79}$$

Determine the expectation values $\langle m\rangle$ and $\langle m^2\rangle$ for this distribution in terms of the mean density $\bar{n}_{\rm g}$ and the volume $v$ of a sub-volume.

**(b)** Calculate the correlation function of galaxies assuming the Poisson model of Eq. (14.79). That is, compute

$$\xi_{\rm g}(|\boldsymbol{x}_\alpha - \boldsymbol{x}_\beta|) = \frac{\langle m(\boldsymbol{x}_\alpha) m(\boldsymbol{x}_\beta)\rangle}{\langle m\rangle^2}, \tag{14.80}$$

where $\boldsymbol{x}_\alpha$ denotes the position of sub-volume $\alpha$. Assume that there is no intrinsic clustering, so that $\langle m\rangle$ is the same in each sub-volume.

**(c)** Compute the galaxy power spectrum. You can either Fourier-transform the result of the previous step, or compute it directly as follows. Rewrite Eq. (14.49) as

$$\Delta(\boldsymbol{k}_i) = L^{3/2} \sum_\alpha e^{i\boldsymbol{k}_i\cdot\boldsymbol{x}_\alpha} \left[\frac{n(\boldsymbol{x}_\alpha) - \bar{n}_{\rm g}}{\bar{n}_{\rm g}}\right] \tag{14.81}$$

where the sum runs over the sub-volumes of size $v$. Using the results of **(a)**, and again assuming that there is no intrinsic clustering, determine $\langle \Delta(\boldsymbol{k}_i)\Delta(\boldsymbol{k}_j)\rangle$. Using either approach, show that you obtain the noise contribution to Eq. (14.57) with $P_N = 1/\bar{n}_{\rm g}$.

**14.9** Estimate the expected error on the $B$-mode polarization power spectrum $\mathrm{Var}[\hat{C}_{BB}(l)]$ from the BICEP2/Keck Array experiment (Ade et al., 2018). The observations cover 400 square degrees on the sky; assume that the noise in an arcminute-squared pixel is 3 μK. Use the Fisher matrix to calculate the expected upper limit on the tensor-to-scalar ratio $r$ that this experiment is expected to achieve (under ideal circumstances) under the assumption that the true $r = 0$.

**14.10** In the derivation of the Fisher matrix, we assumed that, if averaged over many noise realizations, the measured $\hat{C}(l)$ are equal to the theory prediction $C^{\rm theory}(l, \bar{\lambda})$ at true parameter values $\bar{\lambda}$. The Fisher formalism also allows us to infer the bias on parameters $\lambda$ that is caused by a mismatch between theory and data, either due to systematics in the data or an inadequate theory prediction.

**(a)** Assume that the maximum of the likelihood is attained when

$$\hat{C}(l) = C^{\rm theory}(l, \bar{\lambda}) + C^{\rm sys}(l), \tag{14.82}$$

where $C^{\rm sys}(l)$ is the observational systematic or theory error, while the statistical error $\mathrm{Var}[\hat{C}(l)]$ is unchanged. Assume a single parameter $\lambda$ for simplicity, and derive an expression for the value $\lambda_{\rm sys}$ where the likelihood peaks at linear order in $C^{\rm sys}(l)$ (ignore terms that involve two powers of $C^{\rm sys}(l)$). Hint: perform a Taylor expansion of the likelihood around $\bar{\lambda}$.

**(b)** Take the expectation value of all quantities in this expression, recalling that $\langle\hat{C}(l)\rangle = C^{\rm theory}(l,\bar{\lambda}) + C^{\rm sys}(l)$, and use the Fisher matrix.

**(c)** Now generalize this result to multiple parameters.

**14.11** Derive the Fisher matrix for the galaxy power spectrum. Use this together with Eq. (14.60) and Eq. (11.23) to estimate the expected error on the growth rate $f$ as measured from the galaxy power spectrum for the Euclid survey, assuming the following very approximate specifications:

$$V = 63h^{-3}\,\mathrm{Gpc}^3; \quad z = 1.4; \quad \bar{n}_{\rm g} = 5.2\times10^{-4}h^3\,\mathrm{Mpc}^{-3}; \quad b_1 = 1.5. \tag{14.83}$$

You can assume a cubic volume and that $\bar{n}_{\rm g}$ is a fixed parameter, but you should marginalize over $b_1$. Use the results of Sect. 12.2 to propose a sensible value for $k_{\rm max}$ up to which you would trust this forecast based on linear theory (see next exercise).

**14.12** Adapt the Fisher parameter bias formalism derived in Exercise 14.10 from the case of $C(l)$ to the case of the three-dimensional galaxy power spectrum. Use this to estimate the bias on the growth rate $f$ incurred if one were to neglect the 1-loop correction to the matter power spectrum (Eq. (12.48)), as a function of $k_{\rm max}$. At what value of $k_{\rm max}$ does the systematic shift become equal to the 1-$\sigma$ statistical error on $f$?

# A

# Solutions to selected exercises

The exercises at the end of each chapter have a broad range of difficulty. Some are simply repeating calculations in the text in a slightly different context; others are fairly elementary applications of basic formulae; while some are quite challenging. We generally refer to those exercises that are most important for following the main thread of the book in the text of each chapter. This appendix contains solutions to selected exercises.

## Chapter 1

*Exercise 1.1*

The ratio

$$\frac{\rho_\Lambda}{3H^2/(8\pi G)} = \frac{\rho_\Lambda}{\rho_{\rm cr}}\left(\frac{H_0}{H}\right)^2 \tag{A.1}$$

evaluates to 0.7 today. By assumption, the universe is forever radiation dominated (clearly not true today, but a good approximation early on), so we take $H/H_0 = a^{-2}$. The temperature also scales as $a^{-1}$, so $H/H_0 = (T/T_0)^2$ with $T_0 = 2.7\ {\rm K} = 2.3 \times 10^{-4}$ eV. So,

$$\frac{\rho_\Lambda}{3H^2/(8\pi G)} = 0.7\left(\frac{T_0}{T}\right)^4. \tag{A.2}$$

At the Planck scale, $T_0/T = 2.3 \times 10^{-4}/1.22 \times 10^{28}$, so

$$\frac{\rho_\Lambda}{3H^2/(8\pi G)} = 9 \times 10^{-128}. \tag{A.3}$$

This is the so-called fine-tuning problem: for the cosmological constant to be just important today, it had to have been fine-tuned to an absurdly small value at early times; if it had been much larger, our universe would look very different, exponentially expanding and essentially empty by now. It is a deep problem.

*Exercise 1.2*

We need to do the integral

$$t_0 = \frac{1}{H_0}\int_0^1 \frac{da}{a}\left[\Omega_\Lambda + \frac{1-\Omega_\Lambda}{a^3}\right]^{-1/2} \tag{A.4}$$

for $\Omega_\Lambda = 0.7$ and 0. The latter case can be done straightforwardly:

$$\int_0^1 \frac{da}{a} a^{3/2} = \frac{2}{3}. \tag{A.5}$$

So $t_0 = 2/(3H_0) = 0.67 \times 10^{10} h^{-1}$ yr. When $\Omega_\Lambda$ is not zero, the integral can be done using the substitution hinted at in the exercise, or simply numerically. Either way, the result is

$$\int_0^1 \frac{da}{a} \left[0.7 + \frac{0.3}{a^3}\right]^{-1/2} \simeq 0.96. \tag{A.6}$$

So for fixed Hubble constant, a universe with $\Lambda$ is older than a matter-dominated one by a factor of $0.96/0.67 = 1.43$. This is because the universe with $\Lambda$ is accelerating now, so in the past it was expanding more slowly than the matter-dominated one. For $h = 0.7$, a cosmological constant universe has an age of 13.8 billion years, in accord with independent age indicators such as stars and globular clusters.

*Exercise 1.4*

Let us rewrite Eq. (1.9) as

$$I_\nu = \frac{2(2\pi\hbar\nu)\nu^2/c^2}{\exp[2\pi\hbar\nu/k_B T] - 1}. \tag{A.7}$$

The units of the $y$-axis in Fig. 1.7 are million Jansky per steradian, where a Jansky is defined as 1 Jy $= 10^{-26}$ J s$^{-1}$ m$^{-2}$ Hz$^{-1}$; so the units are energy per unit time, area, frequency and solid angle (steradian). On the other hand, $I_\nu$ has units of energy ($2\pi\hbar\nu$) per area ($\nu^2/c^2 =$ m$^{-2}$). In fact, since Hz $=$ s$^{-1}$, these are the same units. So,

$$\text{Intensity [MJy/sr]} = 10^{20}\, I_\nu \text{ [SI units]}. \tag{A.8}$$

Defining $x = 2\pi\hbar\nu/k_B T$, setting the derivative to zero, and solving numerically, we find that Eq. (A.7) attains its maximum when $x \simeq 2.82$. For $T = 2.728$ K, this yields $\nu_{\max} = 160$ GHz. Since $\nu = c/\lambda$, this is also the maximum position in inverse wavelength, and we obtain $(\lambda^{-1})_{\max} = 5.35$ cm$^{-1}$. Plugging in the remaining constants, we find that the peak intensity is

$$I_\nu(\nu_{\max}) = 385 \text{ MJy sr}^{-1}. \tag{A.9}$$

Both of these values match Fig. 1.7.

# Chapter 2

*Exercise 2.1*

**(a)** To get from Kelvin to eV, use $k_B =$ eV/(11605 K). So

$$2.726 \text{ K} \to k_B 2.726 \text{ K} = (2.726/11605) \text{ eV}. \tag{A.10}$$

Or $2.349 \times 10^{-4}$ eV.

**(b)** Since $T_0 = 2.349 \times 10^{-4}$ eV,

$$\rho_\gamma = \frac{\pi^2 T_0^4}{15} = 2.004 \times 10^{-15}\ \text{eV}^4. \tag{A.11}$$

To get this in g cm$^{-3}$, first divide by $(\hbar c)^3 = (1.973 \times 10^{-5}\ \text{eV cm})^3$ to get 0.2609 eV cm$^{-3}$. Then to change from eV to grams, remember that the mass of the proton is either $1.673 \times 10^{-24}$ g or $0.9383 \times 10^9$ eV, so 1 eV $= 1.783 \times 10^{-33}$ g. Therefore, $\rho_\gamma = 4.651 \times 10^{-34}$ g cm$^{-3}$.

**(c)** We have parametrized $H_0 = 100h$ km s$^{-1}$ Mpc$^{-1}$, or using the fact that one Mpc is equal to $3.1 \times 10^{19}$ km, $H_0 = 3.23h \times 10^{-18}$ s$^{-1}$. To get this into inverse cm, divide by the speed of light, $c = 3 \times 10^{10}$ cm s$^{-1}$; then $H_0 = 1.1\,h \times 10^{-28}$ cm. Or $H_0^{-1} = 9.3h^{-1} \times 10^{27}$ cm.

**(d)** To get the Planck mass ($1.2 \times 10^{28}$ eV) into Kelvin, multiply by $k_\text{B}^{-1} = 11605$ K/eV; then $m_\text{Pl} = 1.4 \times 10^{32}$ K. To get it into inverse cm, divide by $\hbar c = 1.97 \times 10^{-5}$ eV cm to get $m_\text{Pl} = 6.1 \times 10^{32}$ cm$^{-1}$. To get this is units of time, multiply by the speed of light to get $m_\text{Pl} = 6.1 \times 10^{32} \times 3 \times 10^{10}$ cm s$^{-1}$, or $m_\text{Pl} = 1.8 \times 10^{43}$ s$^{-1}$.

*Exercise 2.4*

Accumulating the various Christoffel components leads to

$$\frac{d^2x^i}{d\lambda^2} = -2\frac{\dot{a}}{a}\frac{dt}{d\lambda}\frac{dx^i}{d\lambda}. \tag{A.12}$$

Change to differentiation with respect to $\eta$ using the facts that $dt/d\lambda = E$ and $d\eta/d\lambda = E/a$. Then the geodesic equation becomes

$$\frac{E}{a}\frac{d}{d\eta}\left(\frac{E}{a}\frac{dx^i}{d\eta}\right) = -2\frac{\dot{a}}{a}\frac{E^2}{a}\frac{dx^i}{d\eta}. \tag{A.13}$$

Since $E/a \propto a^{-2}$ for massless particles, when the derivative on the left acts on $E/a$, the resulting term (proportional to $dx^i/d\eta$) exactly cancels the term on the right, leaving the result of Eq. (2.94).

*Exercise 2.5*

The age integral is

$$t(a) = \int_0^a \frac{da'}{a'H(a')}. \tag{A.14}$$

Since we are assuming only matter and radiation, we can take

$$H(a) = H_0\sqrt{\rho/\rho_\text{cr}} = H_0\sqrt{\frac{\Omega_\text{m}}{a^3} + \frac{\Omega_\text{r}}{a^4}}. \tag{A.15}$$

This expression is valid at early times when the contribution from the cosmological constant can be neglected. But the ratio $\Omega_\text{r}/\Omega_\text{m} = a_\text{eq} = 4.15 \times 10^{-5}/(\Omega_\text{m}h^2)$. Therefore, the age

integral is

$$t = \frac{1}{\Omega_{\rm m}^{1/2} H_0} \int_0^a \frac{da' a'}{\sqrt{a' + a_{\rm eq}}}. \tag{A.16}$$

Integrate by parts to get

$$\Omega_{\rm m}^{1/2} H_0 t = 2a\sqrt{a + a_{\rm eq}} - 2\int_0^a da' \sqrt{a' + a_{\rm eq}}. \tag{A.17}$$

Carrying out the last integral leads to

$$\Omega_{\rm m}^{1/2} H_0 t = 2a\sqrt{a + a_{\rm eq}} - \frac{4}{3}\left\{[a + a_{\rm eq}]^{3/2} - a_{\rm eq}^{3/2}\right\}. \tag{A.18}$$

At very early times, such as when the temperature was 0.1 MeV, $a$ is much smaller than $a_{\rm eq}$, so

$$t = \frac{a^2}{2H_0\sqrt{\Omega_{\rm m} a_{\rm eq}}} \qquad \left(a \ll a_{\rm eq}\right). \tag{A.19}$$

This limit is easiest to see directly in the integral of Eq. (A.16), but you can also get it by Taylor expanding Eq. (A.18). When the temperature is 0.1 MeV, the scale factor is the temperature today divided by 0.1 MeV, so $2.35 \times 10^{-4}$ eV/0.1 MeV $= 2.35 \times 10^{-9}$. Plugging in numbers leads to

$$t(0.1\ {\rm MeV}) = 4.28 \times 10^{-16} \times 9.78 \times 10^9 h^{-1}\ {\rm yr} = 132\ {\rm s}. \tag{A.20}$$

Note that this result does not depend on the value of $\Omega_{\rm m}$ since radiation dominates at early times. At $T = 1/4$ eV, $a = 9.4 \times 10^{-4}$. Plugging in concordance cosmology parameters, Eq. (A.18) leads to

$$t\ (1/4\,{\rm eV}) = 389,000\ {\rm yr}. \tag{A.21}$$

*Exercise 2.7*

The angle subtended is the physical size divided by the angular diameter distance

$$\theta(z) = 5\ {\rm kpc}\frac{1+z}{\chi(z)}. \tag{A.22}$$

In a Euclidean matter-dominated universe, $\chi$ is given by

$$\begin{aligned}
\chi^{\rm Euclidean,MD}(a) &= \int_a^1 \frac{da'}{H_0^{1/2} a'^{\,1/2}} \\
&= \frac{2}{H_0}\left[1 - a^{1/2}\right] \\
&= \frac{2}{H_0}\left[1 - \frac{1}{\sqrt{1+z}}\right].
\end{aligned} \tag{A.23}$$

When $z = 0.1$ (1), the term in brackets in Eq. (A.23) is equal to 0.0465 (0.293). The comoving distance out to $z$ is, therefore,

$$\chi = \begin{cases} 279h^{-1}\,\text{Mpc} & z = 0.1 \\ 1756h^{-1}\,\text{Mpc} & z = 1 \end{cases} . \tag{A.24}$$

Carrying out the division and converting radians to arcsec (1 radian equals $2.06 \times 10^5$ arcsec) leads to

$$\theta = \begin{cases} 4.07'' h & z = 0.1 \\ 1.17'' h & z = 1 \end{cases} . \tag{A.25}$$

In a universe with $\Omega_\Lambda > 0$, $\chi$ must be computed numerically. At $z = 1$, one finds $\chi$ to be larger than in the Euclidean matter-dominated case by 30% in the fiducial cosmology, so the angular size will be smaller by this factor: $\theta = 0.90'' h$. At $z = 0.1$ the difference in comoving distances is only a few percent, so the angular size goes down to $\theta = 3.88'' h$ in the fiducial cosmology.

*Exercise 2.8*

Rewriting Eq. (1.9) in terms of momentum $p = 2\pi\hbar\nu/c$ and recognizing the denominator there as $1/f_{\rm BE}$ leads to

$$I_\nu = f_{\rm BE}(p)\frac{4\pi p^3}{(2\pi)^3} \tag{A.26}$$

with $\hbar = c = 1$ (in Ch. 13 we will develop the physical content of this relation). The energy density is the integral of this over all frequencies, with a factor of $4\pi$ to count photons from all directions (since $I_\nu$ counts the energy flux per steradian):

$$\rho_\gamma = 4\pi \int_0^\infty d\nu I_\nu . \tag{A.27}$$

This can be converted into an integral over momentum, with $d\nu = dp/(2\pi)$:

$$\rho_\gamma = 2 \int_0^\infty dp I_\nu . \tag{A.28}$$

*Exercise 2.11*

The energy density of a massless boson is $g\pi^2 T^4/30$, while that of a fermion is 7/8 times this. So,

$$s = \frac{2\pi^2}{45}\left[\sum_{i=\text{bosons}} g_i T_i^3 + \frac{7}{8}\sum_{i=\text{fermions}} g_i T_i^3\right] \tag{A.29}$$

accounting for the possibility that different species have different temperatures. For a massive particle with $\mu = 0$, at temperature far below the mass, $e^{E/T} \to e^{m/T} \times e^{p^2/2mT}$. So, for

both fermions and bosons, the distribution function and hence both the pressure and the energy density scale as $e^{-m/T}$.

# Chapter 3

*Exercise 3.2*

**(a)** Start with

$$\Gamma^0{}_{\mu\nu} = \frac{g^{0\alpha}}{2}\left[\frac{\partial g_{\alpha\mu}}{\partial x^\nu} + \frac{\partial g_{\alpha\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\alpha}\right] \tag{A.30}$$

where $\mu$, $\nu$ range from 0 to 2, 0 being the time index, 1 corresponding to $\theta$, and 2 to $\phi$. Since the metric is diagonal, $g^{0\alpha}$ is nonzero only when $\alpha = 0$ in which case it is $-1$. So

$$\Gamma^0{}_{\mu\nu} = -\frac{1}{2}\left[\frac{\partial g_{0\mu}}{\partial x^\nu} + \frac{\partial g_{0\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial t}\right]. \tag{A.31}$$

All of these terms vanish: the first two since $g_{00}$ is a constant, and the last because none of the metric elements depend on $x^0 = t$. So $\Gamma^0{}_{\mu\nu} = 0$ for all $\mu$, $\nu$.
Next consider

$$\Gamma^\theta{}_{\mu\nu} = \frac{g^{\theta\alpha}}{2}\left[\frac{\partial g_{\alpha\mu}}{\partial x^\nu} + \frac{\partial g_{\alpha\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\alpha}\right]. \tag{A.32}$$

Again since the metric is diagonal, and $g^{\theta\theta} = 1/r^2$, this reduces to

$$\Gamma^\theta{}_{\mu\nu} = \frac{1}{2r^2}\left[\frac{\partial g_{\theta\mu}}{\partial x^\nu} + \frac{\partial g_{\theta\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial \theta}\right]. \tag{A.33}$$

Only the $g_{\phi\phi}$ component depends on one of our variables, so only it is nonzero when differentiated. Therefore, the first two terms vanish and the last is nonzero only when $\mu = \nu = \phi$, in which case it is

$$\Gamma^\theta{}_{\phi\phi} = \frac{1}{2r^2}\left[-r^2\frac{\partial \sin^2\theta}{\partial\theta}\right] = -\sin\theta\cos\theta. \tag{A.34}$$

Finally, when the upper index is $\phi$, we have

$$\Gamma^\phi{}_{\mu\nu} = \frac{1}{2r^2\sin^\theta}\left[\frac{\partial g_{\phi\mu}}{\partial x^\nu} + \frac{\partial g_{\phi\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial \phi}\right]. \tag{A.35}$$

The last term vanishes since none of the metric elements depend on $\phi$; the first two are nonzero only if one of the indices $\mu$, $\nu$ is equal to $\phi$ and the other is $\theta$, so

$$\Gamma^\phi{}_{\phi\theta} = \Gamma^\phi{}_{\theta\phi} = \frac{\cos\theta}{\sin\theta}. \tag{A.36}$$

**(b)** The geodesic equation is

$$\frac{d^2x^\mu}{d\lambda^2} = -\Gamma^\mu{}_{\alpha\beta} P^\alpha P^\beta \tag{A.37}$$

with

$$P^\mu \equiv \frac{dx^\mu}{d\lambda}. \tag{A.38}$$

Let us apply this to the $\mu = \theta$ component. The left-hand side is

$$\frac{d^2\theta}{d\lambda^2} = \frac{d}{d\lambda}\frac{dt}{d\lambda}\dot{\theta} = E^2\ddot{\theta} \tag{A.39}$$

since $E = dt/d\lambda$ is constant. The Christoffel symbol on the right-hand side $\Gamma^\theta{}_{\alpha\beta}$ is nonzero only when $\alpha = \beta = \phi$ in which case it is $-\sin\theta\cos\theta$. So,

$$\ddot{\theta} - \sin\theta\cos\theta(\dot{\phi})^2 = 0. \tag{A.40}$$

For the second equation, consider the $\phi$ component of the geodesic equation,

$$\frac{d^2\phi}{d\lambda^2} = -\Gamma^\phi{}_{\alpha\beta} P^\alpha P^\beta. \tag{A.41}$$

Again the left-hand side is simply $E^2\ddot{\phi}$. The right-hand side gets nonzero contributions when $\alpha = \theta$, $\beta = \phi$ or an identical term when $\alpha = \phi$, $\beta = \theta$. Therefore,

$$\ddot{\phi} + 2\frac{\cos\theta}{\sin\theta}\dot{\theta}\dot{\phi} = 0. \tag{A.42}$$

Incidentally, this is equivalent to

$$\frac{d}{dt}\left(\dot{\phi}\sin^2\theta\right) = 0 \tag{A.43}$$

and the conserved quantity in parentheses is the angular momentum.

**(c)** The time-time component of the Ricci tensor $R_{00}$ vanishes since all $\Gamma$ with time components are zero. We need to compute the spatial components. First, consider

$$R_{\theta\theta} = \frac{\partial\Gamma^\alpha{}_{\theta\theta}}{\partial x^\alpha} - \frac{\partial\Gamma^\alpha{}_{\theta\alpha}}{\partial\theta} + \Gamma^\alpha{}_{\beta\alpha}\Gamma^\beta{}_{\theta\theta} - \Gamma^\alpha{}_{\beta\theta}\Gamma^\beta{}_{\theta\alpha}. \tag{A.44}$$

The first and third terms vanish since the Christoffel symbol with two lower $\theta$ vanishes. For the same reason, the index $\alpha$ in the second term must be equal to $\phi$, and both $\beta$ and $\alpha$ in the last term must equal $\phi$:

$$R_{\theta\theta} = -\frac{\partial(\cos\theta/\sin\theta)}{\partial\theta} - \left(\frac{\cos\theta}{\sin\theta}\right)^2. \tag{A.45}$$

Carrying out the derivative then gives

$$R_{\theta\theta} = \left[1 + \frac{\cos^2\theta}{\sin^2\theta}\right] - \left(\frac{\cos\theta}{\sin\theta}\right)^2 = 1. \tag{A.46}$$

The other spatial component is

$$R_{\phi\phi} = \frac{\partial\Gamma^{\alpha}{}_{\phi\phi}}{\partial x^{\alpha}} - \frac{\partial\Gamma^{\alpha}{}_{\phi\alpha}}{\partial\phi} + \Gamma^{\alpha}{}_{\beta\alpha}\Gamma^{\beta}{}_{\phi\phi} - \Gamma^{\alpha}{}_{\beta\phi}\Gamma^{\beta}{}_{\phi\alpha}. \tag{A.47}$$

The Christoffel symbol in the first term is nonzero only if $\alpha = \theta$, while the one in the second term is always zero. In the third term $\beta$ must be equal to $\theta$ to make the second Christoffel symbol be nonzero, and then $\alpha = \phi$. In the last term $\beta$ can be $\theta$ and $\alpha = \phi$ or vice versa, so

$$R_{\phi\phi} = \frac{\partial\Gamma^{\theta}{}_{\phi\phi}}{\partial\theta} + \Gamma^{\phi}{}_{\theta\phi}\Gamma^{\theta}{}_{\phi\phi} - \Gamma^{\phi}{}_{\theta\phi}\Gamma^{\theta}{}_{\phi\phi} - \Gamma^{\theta}{}_{\phi\phi}\Gamma^{\beta}{}_{\phi\theta}. \tag{A.48}$$

The middle two terms cancel leaving

$$R_{\phi\phi} = -\frac{\partial(\sin\theta\cos\theta)}{\partial\theta} + \sin\theta\cos\theta\frac{\cos\theta}{\sin\theta}. \tag{A.49}$$

Carrying out the derivative gives

$$R_{\phi\phi} = -\cos^2\theta + \sin^2\theta + \cos^2\theta = \sin^2\theta. \tag{A.50}$$

Finally, the Ricci scalar is

$$R = g^{\mu\nu}R_{\mu\nu} = -R_{00} + \frac{1}{r^2}R_{\theta\theta} + \frac{1}{r^2\sin^2\theta}R_{\phi\phi}. \tag{A.51}$$

Assembling the terms, we get

$$R = \frac{1}{2r^2}. \tag{A.52}$$

The Ricci scalar is therefore a measure of the curvature of the space.

*Exercise 3.6*

Combining Eq. (2.60) and Eq. (3.90) yields

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\sum_s (1 + 3w_s)\rho_s, \tag{A.53}$$

where the sum runs over all constituents. For a single constituent, the condition for acceleration becomes, simply enough,

$$\frac{\ddot{a}}{a} > 0 \quad \Leftrightarrow \quad w < -\frac{1}{3}, \tag{A.54}$$

since $\rho_s > 0$ always. No ordinary form of matter (relativistic or non-relativistic) has such an equation of state. Neither does curvature.

If there are multiple constituents, we obtain the condition

$$\frac{\sum_s w_s \rho_s}{\sum_s \rho_s} < -\frac{1}{3}. \tag{A.55}$$

That is, the density-weighted mean equation of state has to be less than $-1/3$.

*Exercise 3.7*

**(a)** Using the homogeneous FRW metric, we have

$$P_0 = g_{00} P^0 = -P^0 = -E; \qquad P_i = a^2 P^i = a p^i. \tag{A.56}$$

Since $p^i \propto 1/a$, we see that $P_i$ is constant in the homogeneous universe (it is in fact the "superconformal momentum" used in N-body simulations, Sect. 12.3).

**(b)** Again for the homogeneous universe, we have $\sqrt{-\det[g_{\alpha\beta}]} = a^3$, and

$$\begin{aligned} T^0{}_0(\boldsymbol{x}, t) &= \frac{g}{a^3} \int \frac{dP_1 dP_2 dP_3}{(2\pi)^3} P_0 f(\boldsymbol{p}, t) \\ &= -g \int \frac{d^3 p}{(2\pi)^3} E(p) f(\boldsymbol{p}, t) = -\rho, \end{aligned} \tag{A.57}$$

where $\rho$ is the energy density obtained from the distribution function, Eq. (2.62).

**(c)** One third of the sum of the diagonal spatial components of the stress-energy tensor is given by

$$\begin{aligned} \frac{1}{3} T^k{}_k(\boldsymbol{x}, t) &= \frac{1}{3} \frac{g}{a^3} \int \frac{dP_1 dP_2 dP_3}{(2\pi)^3} \frac{P^k P_k}{P^0} f(\boldsymbol{p}, t) \\ &= \frac{1}{3} g \int \frac{d^3 p}{(2\pi)^3} \frac{p^2}{E(p)} f(\boldsymbol{p}, t) = \mathcal{P}, \end{aligned} \tag{A.58}$$

since $P^k P_k = p^k p_k = p^2$. So we recover Eq. (2.64) as expected.

*Exercise 3.8*

First integrate Eq. (3.23) over all momentum. This gives

$$\frac{\partial n}{\partial t} + \frac{\partial (nu)}{\partial x} = 0, \tag{A.59}$$

the $\partial f/\partial p$ term vanishing after integrating by parts and noticing that $f = 0$ at $p = \pm\infty$ (there are no particles with infinite momentum). This is the continuity equation. To get the Euler equation, first multiply by $p/m$ and then integrate over all momentum. This gives

$$\frac{\partial (nu)}{\partial t} + \frac{\partial}{\partial x} \int_{-\infty}^{\infty} \frac{dp}{2\pi} \frac{p^2}{m^2} f(x, p, t) + \frac{kx}{m} n = 0 \tag{A.60}$$

where the last term follows from an integration by parts. The integral in the second term yields two contributions,

$$\int_{-\infty}^{\infty} \frac{dp}{2\pi} \frac{p^2}{m^2} f(x, p, t) = nu^2(x, t) + \sigma(x, t), \tag{A.61}$$

the first a *bulk velocity* term, and the second the *velocity dispersion* or second moment $\sigma$. Using the continuity equation reduces Eq. (A.60) to

$$\dot{u} + u\frac{\partial u}{\partial x} + \frac{1}{n}\frac{\partial \sigma}{\partial x} + \frac{kx}{m} = 0. \tag{A.62}$$

The second moment acts like a pressure term. In order to close the set of equations, we need to either set $\sigma = 0$ or relate it to the other variables, $n$ and $u$.

*Exercise 3.12*

Eq. (3.60) together with Eq. (3.49) yield

$$P_\mu = \left[-E(1+\Psi),\ p^i a(1+\Phi)\right], \tag{A.63}$$

while, using $1/\sqrt{-\det[g_{\alpha\beta}]} = a^{-3}(1-\Psi-3\Phi)$, Eq. (3.20) becomes

$$T^\mu{}_\nu = g(1-\Psi)\int \frac{d^3p}{(2\pi)^3} \frac{P^\mu P_\nu}{P^0} f(\boldsymbol{x}, \boldsymbol{p}, t). \tag{A.64}$$

Now, plugging in Eq. (3.60), Eq. (A.63), and expanding to first order in $\Phi$, $\Psi$ yields Eq. (3.86).

# Chapter 4

*Exercise 4.1*

The number density of a species with degeneracy $g = 2$ is

$$n = 2\int \frac{d^3p}{(2\pi)^3} f(p). \tag{A.65}$$

For the distributions we will consider, the phase space density $f$ depends only on the magnitude of the momentum, so the angular part of the integral can be performed leading to a factor of $4\pi$; therefore,

$$n = \frac{1}{\pi^2}\int_0^\infty dp\ p^2 f(p). \tag{A.66}$$

First let us consider the low-temperature limit, $m/T \gg 1$. In this case, the limit of the Boltzmann distribution is $\exp[-(m + p^2/2m)/T]$. We claim, though, that this is precisely the limit of both the Fermi–Dirac and Bose–Einstein distributions:

$$\frac{1}{e^{E/T} \pm 1} \to e^{-E/T} \tag{A.67}$$

since $E \simeq m \gg T$ so that the exponential in the denominator dwarfs the 1. Therefore the low-temperature limit of all three distributions is

$$n^{\text{low T}} = \frac{e^{-m/T}}{\pi^2} \int_0^\infty dp\, p^2 e^{-p^2/2mT}. \tag{A.68}$$

To do the integral, define a dimensionless parameter $x \equiv p/\sqrt{2mT}$. Then, $dpp^2 = [2mT]^{3/2} dx x^2$, so

$$n^{\text{low T}} = \frac{e^{-m/T}}{\pi^2} [2mT]^{3/2} \int_0^\infty dx\, x^2 e^{-x^2}. \tag{A.69}$$

But the integral is equal to $\sqrt{\pi}/4$, so we have

$$n^{\text{low T}} = 2e^{-m/T} \left(\frac{mT}{2\pi}\right)^{3/2}. \tag{A.70}$$

The high-temperature limit of the Boltzmann distribution is

$$n^{\text{Hi T, Boltz}} = \frac{1}{\pi^2} \int_0^\infty dp\, p^2 e^{-p/T}. \tag{A.71}$$

Defining the integration variable $x \equiv p/T$ leads to

$$n^{\text{Hi T, Boltz}} = \frac{1}{\pi^2} T^3 \int_0^\infty dx\, x^2 e^{-x}. \tag{A.72}$$

The $x$ integral is equal to 2. So,

$$n^{\text{Hi T, Boltz}} = \frac{2T^3}{\pi^2}. \tag{A.73}$$

The Bose–Einstein and Fermi–Dirac integrals similarly are

$$n^{\text{Hi T, BE/FD}} = \frac{T^3}{\pi^2} \int_0^\infty \frac{dx\, x^2}{e^x \mp 1}. \tag{A.74}$$

The integrals can be written in terms of the Riemann zeta function, via Eq. (C.29). So the integral in Eq. (A.74) with the minus sign—the Bose–Einstein distribution—is $\zeta(3)\Gamma(3) = 2\zeta(3)$. The integral with the plus sign—the Fermi–Dirac distribution—is $3\zeta(3)\Gamma(3)/4 = 3\zeta(3)/2$, so

$$n^{\text{Hi T}} = \frac{\zeta(3)T^3}{\pi^2} \begin{cases} 2 & \text{Bose–Einstein,} \\ 3/2 & \text{Fermi–Dirac.} \end{cases}$$

So there are more bosons than fermions for the same temperature, and, since $\zeta(3) \simeq 1.202$, these bracket the Boltzmann amount. All of course are proportional to $T^3$.

*Exercise 4.6*

The photon number density is 411 $\mathrm{cm}^{-3}$, while the baryon number density is $n_\mathrm{b} = \rho_\mathrm{b}/m_p = \rho_\mathrm{cr}\Omega_\mathrm{b}/m_p$. Plugging in numbers gives

$$n_\mathrm{b} = \Omega_\mathrm{b}\frac{1.879h^2 \times 10^{-29}\ \mathrm{g\ cm}^{-3}}{1.673 \times 10^{-24}\ \mathrm{g}} = 1.12 \times 10^{-5}\Omega_\mathrm{b}h^2\ \mathrm{cm}^{-3}. \tag{A.75}$$

So $\eta_\mathrm{b}$, the ratio of the baryon to the photon number density, is indeed given by Eq. (4.10).

*Exercise 4.9*

To find this ratio, we compute the entropy density $(\mathcal{P} + \rho)/T$ at the two times. In both cases, only relativistic particles contribute to the entropy density significantly so that Eq. (A.29) holds. At high temperatures, the following particles contribute to the energy density: quarks ($g_q = 5 \times 3 \times 2$ for the five least massive types—up, down, strange, charm, bottom—each with three colors and two spin states); anti-quarks ($g_{\bar{q}} = 30$ again); leptons ($g_l = 3 \times 2$ for the three massive types—$e$, $\mu$, $\tau$—each with two spin states, and 3 for the corresponding neutrinos with only one helicity state, see Sect. 2.4.4); anti-leptons ($g_{\bar{l}} = 9$ again); photons (2); and gluons ($g_g = 8 \times 2$ for eight possible colors each with two spin states). This totals up to

$$g_* = 2 + 16 + \frac{7}{8}(30 + 30 + 9 + 9) = 86.25. \tag{A.76}$$

The sixth quark, the top quark, does not contribute because it is too heavy to be around at these temperatures: $m_t \simeq 175$ GeV. The same holds for the $W$, $Z$, and Higgs bosons.

Today entropy comes only from photons and neutrinos. The former contribute 2 to $g_*$; the latter contribute $(7/8) \times 3 \times 2 \times (4/11)^{4/3} = 1.36$, so today $g_* = 3.36$. Since the product $sa^3$ remains constant, we have

$$\left[g_*(aT)^3\right]_{T=10\ \mathrm{GeV}} = \left[g_*(aT)^3\right]_{T_0}. \tag{A.77}$$

Therefore,

$$\frac{(aT)^3|_{T=10\ \mathrm{GeV}}}{(a_0T_0)^3} = \frac{3.36}{86.25} = \frac{1}{26}. \tag{A.78}$$

If we repeat the calculation at temperatures beyond $\sim 200$ GeV, where all Standard Model particles contribute, we obtain $g_* = 103.75$ and the ratio Eq. (A.78) evaluates to $1/31$, a fairly minor difference.

# Chapter 5

*Exercise 5.3*

From Eq. (4.2), the electron distribution function peaks at zero momentum, with a maximum value of $e^{(\mu_e - m_e)/T}$. To relate the chemical potential to the density, recall that $n =$

$e^{\mu_e/T} n^{(0)}$, so in the low-temperature limit (Eq. (4.5)):

$$e^{\mu_e/T} = \frac{n_e}{2}\left(\frac{2\pi}{m_e T}\right)^{3/2} e^{m_e/T}. \tag{A.79}$$

So the maximum value of $f_e$ is $(n_e/2)(2\pi/m_e T)^{3/2}$. The number density of electrons is the same as that of the protons, so from the solution of Exercise 4.6 we have $n_e = 1.12 \times 10^{-5}(\Omega_{\rm b}h^2)$ cm$^{-3}$ today including both ionized and captured electrons. Taking the electron temperature to be equal to the photon temperature today gives $2\pi/m_e T = 2.04 \times 10^{-11}$ cm$^2$. Putting back in the factors of $a$ leads to

$$f_e^{\rm MAX} = 10^{-21}\Omega_{\rm b}h^2 a^{-3/2}. \tag{A.80}$$

This expression holds only for $T \le m_e$, corresponding to $a \gtrsim 4.6 \times 10^{-10}$. So, as long as the temperature is well below the electron mass, $f_e$ is very small.

*Exercise 5.4*

The difference between the amplitude squared we used in the derivation in Sect. 5.2 and the more accurate one given in the exercise is $24\pi\sigma_{\rm T}m_e^2[(\hat{\boldsymbol{p}}\cdot\hat{\boldsymbol{p}}')^2 - 1/3]$. The combination in square brackets is 2/3 times the second Legendre polynomial. Rewrite this using the addition formula of spherical harmonics; then the difference becomes

$$\Delta|\mathcal{M}|^2 = 16\pi\sigma_{\rm T}m_e^2\frac{4\pi}{5}\sum_{m=-2}^{2} Y_{2m}(\hat{\boldsymbol{p}})Y_{2m}^*(\hat{\boldsymbol{p}}'). \tag{A.81}$$

This is the quantity we need to insert into the multiple integral in Eq. (5.16) in place of $\mathcal{M}^2$. When we do this, only the $m = 0$ term will contribute since all other $Y_{2m}(\hat{\boldsymbol{p}}')$ have an azimuthal dependence which integrates to zero. Therefore, the new collision term due to anisotropic Compton scattering is

$$\begin{aligned}\Delta C[f(\boldsymbol{p})] = {} & \frac{\pi^2 n_e\sigma_{\rm T}}{p}\mathcal{P}_2(\mu)\int\frac{d^3p'}{(2\pi)^3 p'}\mathcal{P}_2(\hat{\boldsymbol{p}}'\cdot\hat{\boldsymbol{k}}) \\ & \times\left\{\delta_{\rm D}^{(1)}(p-p') + (\boldsymbol{p}-\boldsymbol{p}')\cdot\boldsymbol{u}_{\rm b}\frac{\partial\delta_{\rm D}^{(1)}(p-p')}{\partial p'}\right\}\{f(\boldsymbol{p}') - f(\boldsymbol{p})\},\end{aligned} \tag{A.82}$$

where we used the fact that $Y_{20} = -\sqrt{5}\mathcal{P}_2/\sqrt{4\pi}$. We now perform the angular integral over $d\Omega'$. The only term which survives this integral (at linear order) is the one proportional to $\delta_{\rm D}^{(1)}(p-p')f(\boldsymbol{p}')$, leaving

$$\begin{aligned}\Delta C[f(\boldsymbol{p})] = {} & -\frac{n_e\sigma_{\rm T}}{2p}\mathcal{P}_2(\mu)\int_0^\infty p'dp'\delta_{\rm D}^{(1)}(p-p')p'\frac{\partial f^{(0)}}{\partial p'} \\ & \times\int_{-1}^{1}\frac{d\mu}{2}\mathcal{P}_2(\mu)\Theta(\mu).\end{aligned} \tag{A.83}$$

The angular integral gives $-\Theta_2$. Then integrating over the Dirac $\delta_{\rm D}^{(1)}$-function yields

$$\Delta C[f(\boldsymbol{p})] = p\frac{\partial f^{(0)}}{\partial p} n_e \sigma_{\rm T} \frac{1}{2}\mathcal{P}_2(\mu)\Theta_2. \tag{A.84}$$

This adds a factor of $-\mathcal{P}_2\Theta_2/2$ inside the square brackets of Eq. (5.22) and explains the corresponding term in Eq. (5.67).

# Chapter 6

*Exercise 6.1*

In Fourier space, $G^L_{,jl} \to -k_j k_l G^L$, so

$$\begin{aligned}\epsilon_{ijk} G_{kl,jl} &\to -k^2 \epsilon_{ijk}(\hat{k}_k\hat{k}_j - \hat{k}_j\hat{k}_k/3)G^L \\ &= -\frac{2}{3}k^2\epsilon_{ijk}\hat{k}_j\hat{k}_k G^L = 0\end{aligned} \tag{A.85}$$

since $\epsilon_{ijk}$ is antisymmetric under interchange of $j$ and $k$ while $\hat{k}_j\hat{k}_k$ is symmetric. $G_{ij}$ is also traceless since $\delta_{ij}(\hat{k}_i\hat{k}_j - \delta_{ij}/3) = 0$.

*Exercise 6.3*

From the transformation law of a scalar field, and the definition of $\delta\phi$ through Eq. (6.7), we have

$$\hat{\phi}(\hat{x}) = \phi(x[\hat{x}]) = \bar{\phi}(\hat{t} - \zeta) + \delta\phi(\hat{t} - \zeta, \hat{\boldsymbol{x}} - \boldsymbol{\nabla}\xi). \tag{A.86}$$

We can immediately drop $\zeta$ and $\xi$ in the arguments of $\delta\phi$, since $\delta\phi$ is already of first order. Expanding $\bar{\phi}$ then yields

$$\hat{\phi}(\hat{x}) = \bar{\phi}(\hat{t}) - \zeta\frac{d\bar{\phi}(\hat{t})}{d\hat{t}} + \delta\phi(\hat{t}, \hat{\boldsymbol{x}}). \tag{A.87}$$

On the other hand, in the $\hat{x}$ coordinate system, Eq. (6.7) is

$$\hat{\phi}(\hat{x}) = \bar{\phi}(\hat{t}) + \delta\phi(\hat{t}, \hat{\boldsymbol{x}}). \tag{A.88}$$

Equating the two relations yields

$$\delta\hat{\phi}(\hat{t}, \hat{\boldsymbol{x}}) = \delta\phi(\hat{t}, \hat{\boldsymbol{x}}) - \frac{d\bar{\phi}(\hat{t})}{d\hat{t}}\zeta(\hat{t}, \hat{\boldsymbol{x}}). \tag{A.89}$$

*Exercise 6.8*

**(a)** By definition,

$$\Gamma^i{}_{jk} = \frac{g^{il}}{2}[g_{lj,k} + g_{lk,j} - g_{jk,l}]. \tag{A.90}$$

All derivatives here are spatial, and the only spatially varying part of the metric is the first-order piece $h^{\rm TT}$. Therefore, we can again use the zeroth-order $g^{il} = \delta_{il}/a^2$, leaving Eq. (6.57).

**(b)** The product $\Gamma^{\alpha}{}_{\beta j}\Gamma^{\beta}{}_{i\alpha}$ vanishes when both indices $\alpha$ and $\beta$ are zero (because $\Gamma^{0}{}_{0i} = 0$) and when both indices are spatial (because then each Christoffel symbol is of first order). Therefore, this product is

$$\begin{aligned}\Gamma^{\alpha}{}_{\beta j}\Gamma^{\beta}{}_{i\alpha} &= \Gamma^{0}{}_{kj}\Gamma^{k}{}_{i0} + \Gamma^{k}{}_{0j}\Gamma^{0}{}_{ik} \\ &= \Gamma^{0}{}_{kj}\Gamma^{k}{}_{i0} + (i \leftrightarrow j).\end{aligned} \tag{A.91}$$

But

$$\begin{aligned}\Gamma^{0}{}_{kj}\Gamma^{k}{}_{i0} &= \frac{1}{2}\left(2Hg_{jk} + a^2 h^{\rm TT}_{jk,0}\right)\left(H\delta_{ik} + \frac{1}{2}h^{\rm TT}_{ik,0}\right) \\ &= H^2 g_{ij} + a\dot{a}h^{\rm TT}_{ij,0}.\end{aligned} \tag{A.92}$$

We must remember to add back in the same set of terms with $i$ and $j$ interchanged. This just introduces a factor of 2, so

$$\Gamma^{\alpha}{}_{\beta j}\Gamma^{\beta}{}_{i\alpha} = 2H^2 g_{ij} + 2a\dot{a}h^{\rm TT}_{ij,0}. \tag{A.93}$$

# Chapter 7

## *Exercise 7.2*

There are 411 photons per $\mathrm{cm}^{-3}$ today; the Hubble volume is $(4\pi/3)[3000\,h^{-1}\,\mathrm{Mpc}]^3 = 3.3\times 10^{84}\,h^{-3}\,\mathrm{cm}^3$. So the total number of photons is $1.4\times 10^{87}\,h^{-3}$. This number remains roughly constant throughout the matter and radiation eras since the number density scales as $T^3$, the physical volume as $a^3$, and the temperature as $a^{-1}$. A similar amount of entropy is added by neutrinos. So another problem of the classical cosmology is: Why is the entropy of the universe so large?

The production of entropy actually takes place at the end of inflation during the reheating process: even though the temperature at the end of inflation is extremely small (since any radiation has been extremely diluted by expansion, see Fig. 7.4), the energy density (which is completely in the scalar field) is enormous. When the energy in the scalar field transforms into radiation, the temperature of the radiation shoots up from its value of essentially zero to $T \sim \rho^{1/4}$, which as we have seen can be as high as $10^{14}$ GeV. Thus, the reheating process is responsible for the large entropy we see today. Another way to say this is to point out that inflation is a very ordered state: the universe supercools while the field is away from its true vacuum state, i.e. the minimum of the potential. The transition to the true vacuum is a transition to the very disordered state of equilibrium.

*Exercise 7.12*

**(a)** With this substitution, the equation becomes

$$\frac{d^2\tilde{v}}{d\eta^2} + \frac{2}{\eta}\frac{d\tilde{v}}{d\eta} + \left(k^2 - \frac{2}{\eta^2}\right)\tilde{v}. \tag{A.94}$$

Defining $x \equiv k\eta$, we see that $\tilde{v}$ satisfies the spherical Bessel equation of order 1 (Eq. (C.13)).

**(b)** The two solutions to Eq. (C.13) are $j_1(x)$ and $y_1(x)$. The general solution is therefore $Aj_1 + By_1$. Writing these out explicitly leads to

$$\begin{aligned} v = \eta\tilde{v} &= \eta\left(A\frac{\sin x - x\cos x}{x^2} - B\frac{\cos x + x\sin x}{x^2}\right) \\ &= \frac{1}{2k^2\eta}\Big(e^{ik\eta}[-iA - Ak\eta - B + iBk\eta] \\ &\qquad + e^{-ik\eta}[iA - Ak\eta - B - iBk\eta]\Big). \end{aligned} \tag{A.95}$$

When $k\eta$ is very large and negative, we want $v \to e^{-ik\eta}/\sqrt{2k}$, so the coefficient of $e^{+ik\eta}$ in this limit, proportional to $-A + iB$, must vanish. Thus, $A = iB$. The coefficient of $e^{-ik\eta}$ is

$$\frac{1}{2k^2\eta}\left[-2Ak\eta\right] = -\frac{A}{k}. \tag{A.96}$$

This must equal $(2k)^{-1/2}$, so $A = -(k/2)^{1/2}$. Therefore the correct solution is

$$v = \frac{e^{-ik\eta}}{\sqrt{2k}}\left[1 - \frac{i}{k\eta}\right], \tag{A.97}$$

in agreement with Eq. (7.40).

*Exercise 7.13*

The two components of Einstein's equations are

$$\begin{aligned} k^2\Psi + 3aH\left(\Psi' + aH\Psi\right) &= 4\pi G a^2 \delta T^0{}_0 \\ ik_i(\Psi' + aH\Psi) &= -4\pi G a \delta T^0{}_i. \end{aligned} \tag{A.98}$$

Here we have simply copied the results from Ch. 6, replacing $\Phi$ with $-\Psi$. Since $\delta T^0{}_i$ is first order, we can raise and lower indices with the background metric to obtain $\delta T^i{}_0 = -\delta T^0{}_i/a^2$. Then, multiply the second of these equations by $3iaHk_i/k^2$, and add the two equations to get

$$k^2\Psi = 4\pi G a^2\left[\delta T^0{}_0 + \frac{3a^2 H i k_i \delta T^i{}_0}{k^2}\right]. \tag{A.99}$$

On large scales, the left-hand side is negligible, so the terms in brackets on the right must sum to zero, giving Eq. (7.66).

# Chapter 8

## *Exercise 8.4*

**(a)–(b)** are straightforward.

**(c)** To do the integral, introduce a new integration variable $x \equiv \sqrt{1+y}$. Then Eq. (8.30) becomes

$$\Phi = \frac{3\Phi(0)}{2}\frac{\sqrt{1+y}}{y^3}\int_1^{\sqrt{1+y}} dx \frac{(x^2-1)^2(3x^2+1)}{x^2}. \tag{A.100}$$

Now integrate by parts using the fact that the integral of $1/x^2$ is equal to $-1/x$. The surface term is proportional to the numerator and so vanishes at the lower limit, when $x = 1$. Therefore,

$$\begin{aligned}\Phi &= \frac{3\Phi(0)}{2}\frac{\sqrt{1+y}}{y^3}\left[-\frac{y^2(4+3y)}{\sqrt{1+y}} + \int_1^{\sqrt{1+y}} dx\,(18x^4 - 20x^2 + 2)\right] \\ &= \frac{3\Phi(0)}{2}\frac{\sqrt{1+y}}{y^3}\left[-\frac{y^2(4+3y)}{\sqrt{1+y}} + \left(\frac{18}{5}x^5 - \frac{20}{3}x^3 + 2x\right)\Big|_1^{\sqrt{1+y}}\right].\end{aligned} \tag{A.101}$$

Evaluating the terms in parentheses at the upper and lower limits leads to Eq. (8.31).

## *Exercise 8.8*

**(a)** Let us make the ansatz $\delta_\mathrm{m} = \text{const} \times H$. Since Eq. (8.75) is homogeneous, we can simply take $\delta_\mathrm{m} = H$. We transform the equation to an equation in terms of $H^2$ and $\ln a$, using that

$$\frac{d^2(H^2)}{da^2} = 2H\left(\frac{d\ln H}{da}\frac{dH}{da} + \frac{d^2H}{da^2}\right). \tag{A.102}$$

This yields

$$\frac{d^2(H^2)}{d(\ln a)^2} + 2\frac{d(H^2)}{d\ln a} = \frac{3\Omega_\mathrm{m}H_0^2}{a^3}. \tag{A.103}$$

Now suppose there are several ingredients in the universe, so

$$H^2(a) = H_0^2\sum_s \Omega_s a^{p_s} \qquad \text{where} \qquad p_s = -3(1+w_s). \tag{A.104}$$

Inserting into the equation for $H^2$ yields

$$\sum_s (p_s^2 + 2p_s)\Omega_s a^{p_s} = 3\Omega_\mathrm{m}a^{-3}. \tag{A.105}$$

It is easy to see that the matter component ($p_\mathrm{m} = -3$) fulfills this equation. So for the solution to be valid, any *other* component has to yield a vanishing contribution to the

left-hand side, so

$$\text{not matter:} \qquad p_s^2 + 2p_s = 0 \quad \Leftrightarrow \quad p_s(p_s+2) = 0. \tag{A.106}$$

So, $H$ is a solution to the growth equation in the presence of matter, a cosmological constant ($p_\Lambda = 0$, first solution of the above equation), and curvature ($p_\mathrm{K} = -2$, second solution), but not other possible components.
In all cases, $H$ is a decreasing function of time, so this decaying solution of the growth equation is not what we are after. It does allow us to find the growing solution we are interested in by trying the ansatz $u = \delta_\mathrm{m}/H$. However, the same conditions on the validity we found above also hold for this second solution.

**(b)** We use as ansatz $u = \delta_\mathrm{m}/H$. The evolution equation for $u$ then becomes

$$\frac{d^2u}{da^2} + 3\left[\frac{d\ln(H)}{da} + \frac{1}{a}\right]\frac{du}{da} = 0. \tag{A.107}$$

This first-order equation for $du/dy$ can be integrated to obtain

$$\frac{du}{da} \propto (aH)^{-3}. \tag{A.108}$$

Integrating again and remembering that the growth factor is $uH$ leads to an expression for the growth factor

$$D_+(a) \propto H(a)\int^a \frac{da'}{(a'H(a'))^3}, \tag{A.109}$$

which is the result Eq. (8.77).

**(c)** We found under **(a)** that the solutions we obtained only apply to the case of matter, cosmological constant, and curvature; they do not hold if the dark energy equation of state $w \neq -1$. Fig. A.1 shows the result of the approximate integral Eq. (A.109) in addition to the solution of the differential equation. There is clear disagreement for $w \neq -1$.

*Exercise 8.13*

**(a)** We have

$$\begin{aligned}\sigma_R^2 &= \left\langle\left[\int d^3x\, \delta(\boldsymbol{x}) W_R(|\boldsymbol{x}|)\right]^2\right\rangle \\ &= \left\langle\left[\frac{d^3k}{(2\pi)^3}\delta(\boldsymbol{k}) W_R^*(\boldsymbol{k})\right]^2\right\rangle\end{aligned} \tag{A.110}$$

where we have used the fact that, since $W_R(x)$ is real, $W_R(-\boldsymbol{k}) = W_R^*(\boldsymbol{k})$. Also, we have evaluated $\delta_R$ at the origin. The angular brackets denote the ensemble average

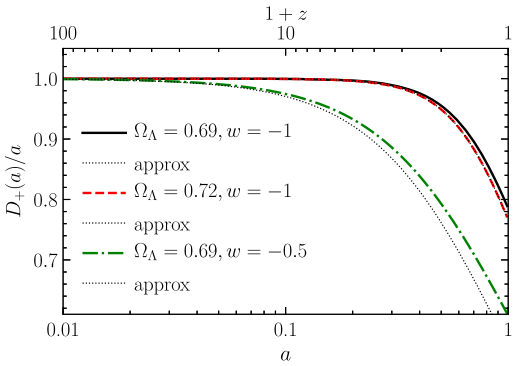



**FIGURE A.1** Same as Fig. 8.15, but also showing the result of the integral solution Eq. (A.109) (dotted). There is clear disagreement if $w \neq -1$.

over all realizations of $\delta(\boldsymbol{k})$. Using Eq. (C.22),

$$\left\langle \delta(\boldsymbol{k})\delta(\boldsymbol{k}')\right\rangle = (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}')\, P_{\rm L}(k) \tag{A.111}$$

leads to

$$\sigma_R^2 = \int \frac{d^3k}{(2\pi)^3} P_{\rm L}(k)\, |W_R(\boldsymbol{k})|^2 \,. \tag{A.112}$$

It remains only to compute the Fourier transform of the tophat window function,

$$\begin{aligned} W_R(\boldsymbol{k}) &= \int d^3x\, W_R(x) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \\ &= \frac{2\pi}{V_R} \int_0^R dx x^2 \int_{-1}^1 d\mu e^{ikx\mu}. \end{aligned} \tag{A.113}$$

Note that we have normalized the window function so that the integral over it is unity; hence the factor of $V_R = 4\pi R^3/3$. Carrying out the remaining angular and radial integrals leads to

$$\begin{aligned} W_R(k) &= \frac{3}{kR^3} \int_0^R dx x \sin(kx) \\ &= \frac{3}{k^3R^3}[-kR\cos(kR) + \sin(kR)]. \end{aligned} \tag{A.114}$$

**(b)–(c)** Evaluating Eq. (A.112) for $R = 8\,h^{-1}$ Mpc, we obtain $\sigma_8 = 0.81$ at $z = 0$ in the fiducial cosmology. $\sigma_R$ is shown as a function of $R$ in Fig. 12.1.

# Chapter 9

## *Exercise 9.2*

Assume a solution of the form $x = e^{i\omega t}$. The damping equation then becomes a quadratic equation for $\omega$:

$$\omega^2 - \frac{ib}{m}\omega - \frac{k}{m} = 0. \tag{A.115}$$

Solving with $k/m > \gamma^2 \equiv (b/2m)^2$ leads to

$$\omega = i\gamma \pm \omega_1. \tag{A.116}$$

The frequency is now $\omega_1 \equiv [k/m - \gamma^2]^{1/2}$, smaller than in the undamped case. The amplitude is also exponentially damped by a factor $e^{-\gamma t}$.

## *Exercise 9.9*

Use the addition theorem of spherical harmonics Eq. (C.12) to write

$$\mathcal{P}_{l'}(\hat{\boldsymbol{p}} \cdot \hat{\boldsymbol{k}}) = \frac{4\pi}{2l+1} \sum_{m'} Y^*_{l'm'}(\hat{\boldsymbol{p}}) Y_{l'm'}(\hat{\boldsymbol{k}}). \tag{A.117}$$

Then the angular integral becomes an integral over the product of two spherical harmonics, which—because of orthogonality—is equal to 1 if $l' = l$ and $m' = m$ and zero otherwise. This leads directly to the desired result.

## *Exercise 9.16*

The generalization of Eq. (9.73) to tensors gives

$$C^{\rm T}(l) = \sum_{l'l''} (-i)^{l'+l''} (2l'+1)(2l''+1) \int \frac{d^3k}{(2\pi)^3} \Theta^{\rm T}_{l'}(k) \Theta^{\rm T*}_{l''}(k) I_{lml'}(k) I^*_{lml''}(k) \tag{A.118}$$

where

$$I_{lml'}(k) \equiv \sqrt{\frac{8\pi}{15}} \int d\Omega\, \mathcal{P}_{l'}(\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{p}}) {\rm Y}_{lm}(\hat{\boldsymbol{p}}) [Y_{22}(\hat{\boldsymbol{p}}) + Y_{2-2}(\hat{\boldsymbol{p}})]. \tag{A.119}$$

The factor of $(8\pi/15)^{1/2}[Y_{22} + Y_{2-2}]$ is the combination $\sin^2\theta\cos(2\phi)$ which appears in Eq. (6.85), so this expression is valid only for the + mode. However, the × mode gives exactly the same result.

The integral $I_{lml'}$ is not trivial. By rewriting the Legendre polynomial as $[4\pi/(2l'+1)]^{1/2} Y_{l'0}/i^{l'}$, we can turn $I_{lml'}$ into an integral over the product of three spherical harmonics. Such integrals are intensively studied in quantum mechanics and can be expressed in terms of the Wigner $3-j$ symbols (see Landau et al., 1965). The integral is then

$$I_{lml'} = \sqrt{\frac{32\pi^2}{15(2l'+1)}} \frac{1}{i^{l'}} \langle lm | Y_{22} + Y_{2-2} | l'0 \rangle \tag{A.120}$$

which vanishes unless $m = 2$ or $m = -2$. When $m$ takes on one of these two values, the matrix element is

$$\langle l2|Y_{22} + Y_{2-2}|l'0\rangle = i^{l'-l} \begin{pmatrix} l & 2 & l' \\ 0 & 0 & 0 \end{pmatrix} \left[\frac{5(2l'+1)(2l+1)}{4\pi}\right]^{1/2} \begin{pmatrix} l & 2 & l' \\ -2 & 2 & 0 \end{pmatrix}. \tag{A.121}$$

The first $3-j$ symbol here, the one with the bottom row all zero, vanishes unless the sum of the elements in the top row $l + l' + 2$ is even. And of course $l'$ cannot differ from $l$ by more than 2 since the combination of $Y_{22}Y_{l'0}$ leads to angular momenta ranging from $l' - 2$ to $l' + 2$. So the only time the matrix element is nonzero is when $l' = l - 2, l, l + 2$. Using Table 9 in Section 106 of Landau et al. (1965) leads to the final result:

$$I_{lml'} = \sqrt{\frac{8\pi}{3}(2l+1)}\, i^{-l}(\delta_{m,2} + \delta_{m,-2})[c_{-2}\delta_{l',l-2} + c_0\delta_{l',l} + c_2\delta_{l',l+2}] \tag{A.122}$$

where here $\delta_{m,2}$ is the Kronecker delta, equal to 1 if $m = 2$ and zero otherwise, and analogously for all other $\delta$. The coefficients are

$$\begin{aligned} c_{-2} &= \frac{\sqrt{6}}{4}\frac{[(l-1)l(l+1)(l+2)]^{1/2}}{(2l-3)(2l-1)(2l+1)} \\ c_0 &= \frac{-2\sqrt{6}}{4}\frac{[(l-1)l(l+1)(l+2)]^{1/2}}{(2l-1)(2l+1)(2l+3)} \\ c_2 &= \frac{\sqrt{6}}{4}\frac{[(l-1)l(l+1)(l+2)]^{1/2}}{(2l+1)(2l+3)(2l+5)}. \end{aligned} \tag{A.123}$$

The result in Eq. (9.94) then follows.

*Exercise 9.17*

**(a)** On large scales, we can take the matter-dominated solution for $h_t$, so

$$\Theta^{\rm T}_{l,t}(k,\eta_0) = -\frac{1}{2}\int_{\eta_*}^{\eta_0} d\eta\; j_l[k(\eta_0-\eta)]\frac{d}{d\eta}\left[\frac{3j_1(k\eta)}{k\eta}\right] h_t(\boldsymbol{k},0). \tag{A.124}$$

Plug this into Eq. (9.94), and use the definition of $P_h(k) = P_{\rm T}(k)/4$ as the power spectrum of superhorizon tensor modes (i.e. for $\eta = 0$) to get

$$\begin{aligned} C^{\rm T}(l) = \frac{1}{2}\frac{9(l-1)l(l+1)(l+2)}{4\pi}\int_0^\infty dk\, k^2 P_{\rm T}(k)\left|\int_0^{\eta_0} d(k\eta)\frac{j_2(k\eta)}{k\eta}\right. \\ \left.\times\left[\frac{j_{l-2}(k[\eta_0-\eta])}{(2l-1)(2l+1)} + 2\frac{j_l(k[\eta_0-\eta])}{(2l-1)(2l+3)} + \frac{j_{l+2}(k[\eta_0-\eta])}{(2l+1)(2l+3)}\right]\right|^2, \end{aligned} \tag{A.125}$$

where we have set the lower limit on the time integral to zero since $\eta_* \ll \eta_0$. Also, we have used the identity $(j_1/x)' = -j_2/x$. The factor of $1/2$ out in front comes from the sum over the $+$ and $\times$ components combined with the conversion to $P_{\rm T}$. Using

Eq. (7.102) for $P_{\rm T}$ (with $n_{\rm T}=0$) and defining new integration variables $y\equiv k\eta_0$ and $x\equiv k\eta$ leads to

$$C^{\rm T}(l)=\frac{9\pi}{2}(l-1)l(l+1)(l+2)\mathcal{A}_{\rm T}\int_0^\infty\frac{dy}{y}\left|\int_0^y dx\frac{j_2(x)}{x}\left[\frac{j_{l-2}(y-x)}{(2l-1)(2l+1)}+2\frac{j_l(y-x)}{(2l-1)(2l+3)}+\frac{j_{l+2}(y-x)}{(2l+1)(2l+3)}\right]\right|^2. \quad \text{(A.126)}$$

**(b)** For the $l=2$ mode, the double integral in Eq. (A.126) is equal to $2.14\times10^{-4}$, so $C^{\rm T}(l=2)=0.036\mathcal{A}_{\rm T}$. The scalar $C(l=2)=4\mathcal{A}_s/75$ in the Sachs–Wolfe limit. This leads to

$$r_2\equiv\frac{C^{\rm T}(l=2)}{C(l=2)}=0.68\frac{\mathcal{A}_{\rm T}}{\mathcal{A}_s}=0.68r. \quad \text{(A.127)}$$

# Chapter 10

*Exercise 10.1*

Under a rotation around the line of sight (the direction of photon propagation), the tensor $I_{ij}$ transforms as

$$\tilde{I}_{ij}=R_i{}^kR_j{}^lI_{kl}, \quad \text{(A.128)}$$

or, in matrix notation, $\tilde{\boldsymbol{I}}=\boldsymbol{R}\boldsymbol{I}\boldsymbol{R}^\top$, where $\boldsymbol{R}$ is the rotation matrix:

$$\boldsymbol{R}(\alpha)=\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}. \quad \text{(A.129)}$$

We obtain

$$\tilde{\boldsymbol{I}}=\begin{pmatrix}I+Q\cos2\alpha-U\sin2\alpha & U\cos2\alpha+Q\sin2\alpha\\ U\cos2\alpha+Q\sin2\alpha & I-Q\cos2\alpha+U\sin2\alpha\end{pmatrix}, \quad \text{(A.130)}$$

from which we can read off the transformation of $I$, $Q$, $U$:

$$\begin{aligned}\tilde{I}&=I,\\ \tilde{Q}&=\cos2\alpha\,Q-\sin2\alpha\,U,\\ \tilde{U}&=\cos2\alpha\,U+\sin2\alpha\,Q.\end{aligned} \quad \text{(A.131)}$$

That is, $Q$ and $U$ depend on the coordinates chosen. Now, if we write $\boldsymbol{l}=l(\cos\phi_l,\sin\phi_l)$, then we have

$$\tilde{\boldsymbol{l}}=\boldsymbol{R}\boldsymbol{l}=l\Big(\cos(\phi_l+\alpha),\sin(\phi_l+\alpha)\Big), \quad \text{(A.132)}$$

so $\tilde{\phi}_l=\phi_l+\alpha$, which comes as no surprise. Plugging $\tilde{Q}$, $\tilde{U}$ as well as $\tilde{\phi}_l$ into Eq. (10.6) and Eq. (10.9), and using the trigonometric relations, we find

$$\tilde{E}=E\quad\text{and}\quad\tilde{B}=B, \quad \text{(A.133)}$$

so $E$ and $B$ are invariant under the coordinate change.

In three dimensions, a parity transformation sends a vector $\boldsymbol{r}$ to $-\boldsymbol{r}$. So, the direction of the line of sight flips; if it came out of the page in Fig. 10.2, it now goes into the page. Rotating it back to the direction out of the page, we see that we have essentially flipped the direction of the $x$-axis (or equivalently only the $y$-axis). Fig. 10.2 shows that we then have

$$\text{parity:} \qquad Q \to Q; \quad U \to -U. \tag{A.134}$$

Further, you can convince yourself that $\phi_l \to \pi - \phi_l$, which implies $\cos 2\phi_l \to \cos 2\phi_l$ while $\sin 2\phi_l \to -\sin 2\phi_l$. Eq. (10.6) and Eq. (10.9) then yield

$$\text{parity:} \qquad E \to E; \quad B \to -B. \tag{A.135}$$

Combining all the results of this exercise, we find that $E$ is *scalar*, while $B$ is *pseudo-scalar*.

*Exercise 10.6*

The angular dependence of the temperature anisotropy induced by a tensor perturbation $h_+$ is

$$\sin^2\theta' \cos(2\phi') = \hat{n}'_x \hat{n}'_x - \hat{n}'_y \hat{n}'_y \tag{A.136}$$

if $\boldsymbol{k}$ is lying along the $z$-axis. Performing the rotation in Eq. (10.49) in order to rotate $\boldsymbol{k}$ to Eq. (10.47), the angular dependence changes to

$$(\cos\alpha\, \hat{n}'_x - \sin\alpha\, \hat{n}'_z)^2 - \hat{n}'_y \hat{n}'_y. \tag{A.137}$$

The $\phi'$ dependences appearing here are $1$, $\cos\phi'$, $\cos^2\phi'$, $\sin^2\phi'$. All of these angular dependences lead to zero when integrating over $\phi'$ weighted by $\sin 2\phi'$ as in Eq. (10.20). Thus, for this particular configuration, $h_+$ does not produce $U$ and hence no $B$-mode polarization. This is a consequence of the fact that the spacetime distortion induced by $h_+$ is aligned with the coordinate axes perpendicular to $\boldsymbol{k}$ (Fig. 6.1). In order to see $B$-mode polarization from $h_+$, we would need to consider a $\boldsymbol{k}$ that is not in the $x-z$-plane.

# Chapter 11

*Exercise 11.1*

Performing two inverse Fourier transforms on Eq. (11.64), we have

$$\begin{aligned}\langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\rangle &= \int \frac{d^3k}{(2\pi)^3} \int \frac{d^3k'}{(2\pi)^3} e^{i(\boldsymbol{x}\cdot\boldsymbol{k}+(\boldsymbol{x}+\boldsymbol{r})\cdot\boldsymbol{k}')} \langle \delta(\boldsymbol{k})\delta(\boldsymbol{k}')\rangle \\ &= \int \frac{d^3k}{(2\pi)^3} e^{-i\boldsymbol{r}\cdot\boldsymbol{k}} P_{\rm L}(k),\end{aligned} \tag{A.138}$$

where the second line follows from Eq. (C.22). This is the (inverse) Fourier transform of the power spectrum. Notice that so far this relation also holds for an anisotropic power spectrum, e.g. in the case of $\delta_{\rm g,obs}$. In the isotropic case, we can perform the angular integrals

to obtain

$$\xi(r) = \frac{1}{2\pi^2}\int_0^\infty k^2 dk \frac{\sin kr}{kr} P_{\rm L}(k). \tag{A.139}$$

*Exercise 11.4*

Eq. (11.66) follows straightforwardly from the orthogonality relation of the Legendre polynomials. Performing the $\mu$ integrals over Eq. (11.23), we obtain

$$P^{(0)}_{\rm g,obs}(k) = \left[1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right] b_1^2 P_{\rm L}(k) \tag{A.140}$$

$$P^{(2)}_{\rm g,obs}(k) = \left[\frac{4}{3}\beta + \frac{4}{7}\beta^2\right] b_1^2 P_{\rm L}(k), \tag{A.141}$$

where $\beta = f/b_1$.

*Exercise 11.8*

To express the angular power spectrum, let us call it $\hat{C}_{\rm g}(l)$, in terms of $w_{\rm g}$, multiply both sides of Eq. (11.69) by $\mathcal{P}_{l'}(\cos\theta)$ and integrate over $\cos\theta$. This gives

$$\hat{C}_{\rm g}(l) = 2\pi \int_{-1}^{1} d\cos\theta\, \mathcal{P}_l(\cos\theta) w_{\rm g}(\theta). \tag{A.142}$$

Express $w_{\rm g}$ as an integral over the 2D power spectrum as in the first line of Eq. (11.49). Then,

$$\hat{C}_{\rm g}(l) = \int_0^\infty dl'\, l' C_{\rm g}(l') \int_{-1}^{1} d\cos\theta\, \mathcal{P}_l(\cos\theta) J_0(l'\theta). \tag{A.143}$$

In the limit that $l'$ is large, the Bessel function becomes

$$J_0(l'\theta) \xrightarrow{l' \gg 1} \mathcal{P}_{l'}(\cos\theta). \tag{A.144}$$

Therefore, the integral over $\theta$ vanishes unless $l = l'$, in which case it is equal to $2/(2l+1)$. The integral over $l'$ is identical to a sum over $l'$ at large $l'$. The factor of $2/(2l+1)$ in the denominator cancels the factor of $l'$ in the numerator, leaving the desired equality between $\hat{C}_{\rm g}(l)$ and $C_{\rm g}(l)$.

# Chapter 12

*Exercise 12.4*

Transform Eq. (8.75) to $x \equiv \ln a$ as time coordinate. Using that $dD_+/dx = fD$, the equation becomes

$$\frac{D_+}{a^2}\left[\frac{df}{dx} + f^2 + \left(\frac{d\ln aH}{dx} + 1\right) f - \frac{3}{2}\frac{\Omega_{\rm m}(\eta_0) H_0^2}{a^3 H^2}\right] = 0. \tag{A.145}$$

Now, by definition of $\Omega_{\rm m}(\eta)$,

$$\Omega_{\rm m}(\eta) = \frac{\rho_{\rm m}(\eta)}{\rho_{\rm cr}(\eta)} = \frac{\Omega_{\rm m}(\eta_0)\rho_{\rm cr}(\eta_0)a^{-3}}{\rho_{\rm cr}(\eta)} = \frac{\Omega_{\rm m}(\eta_0)H_0^2}{a^3H^2}, \tag{A.146}$$

so that the last term in Eq. (A.145) simply becomes $-3\Omega_{\rm m}(\eta)/2$. Finally, we can collect terms and use $d/dx = (aH)^{-1}d/d\eta$ to obtain Eq. (12.32).

Transforming the equation for $\delta^{(2)}$ from Eq. (12.31) into Eq. (12.33) is trivial. For the $\theta_{\rm m}$ equation, we use

$$\theta_{\rm m}{}' = \left(aHf\hat{\theta}\right)' = (aH)^2\left[\frac{3}{2}\Omega_{\rm m}(\eta) - f - f^2\right]\hat{\theta} + (aHf)^2\frac{d\hat{\theta}}{d\ln D_+}. \tag{A.147}$$

Notice that this transformation holds at any order in perturbations, which is why we have dropped the superscript $^{(2)}$. The term $-f(aH)^2\hat{\theta}$ precisely cancels the second term on the left-hand side of the $\theta_{\rm m}$ equation. Finally, dividing both sides by $(aHf)^2$, we obtain the desired source term $\propto D_+^2$ on the right-hand side, while the left-hand side is

$$\frac{d\hat{\theta}}{d\ln D_+} + \left[\frac{3}{2}\frac{\Omega_{\rm m}(\eta)}{f^2} - 1\right]\hat{\theta} + \frac{3}{2}\frac{\Omega_{\rm m}(\eta)}{f^2}\delta, \tag{A.148}$$

which is the left-hand side of the second line in Eq. (12.33).

*Exercise 12.10*

**(a)** The variance of the smoothed density field is given by $\left\langle[\delta_R^{(1)}(\boldsymbol{x})]^2\right\rangle = \sigma(R)$. Thus, $\nu(\boldsymbol{x})$ is a Gaussian random field with mean zero and unit variance, so its probability distribution is given by

$$p(\nu) = \frac{1}{\sqrt{2\pi}}e^{-\nu^2/2}. \tag{A.149}$$

The correlation of the smoothed density field at two different locations is given by the smoothed correlation function, so

$$\left\langle\delta_R^{(1)}(\boldsymbol{x}_1)\delta_R^{(1)}(\boldsymbol{x}_2)\right\rangle = \xi_R(|\boldsymbol{x}_2 - \boldsymbol{x}_1|), \quad \text{so that} \quad \langle\nu_1\nu_2\rangle = \hat{\xi}(r) \equiv \frac{\xi_R(r)}{\sigma^2(R)}, \tag{A.150}$$

where $r = |\boldsymbol{x}_1 - \boldsymbol{x}_2|$. Hence, the joint distribution of $\nu_1$, $\nu_2$ is a bivariate Gaussian with covariance

$$\mathrm{C} = \begin{pmatrix} 1 & \hat{\xi}(r) \\ \hat{\xi}(r) & 1 \end{pmatrix}, \tag{A.151}$$

as given in Eq. (12.111).

**(b)** A simple change of variables $u = \nu/\sqrt{2}$ yields

$$p(\delta_R^{(1)} > \delta_{\rm cr}) = \frac{1}{\sqrt{2\pi}} \int_{\nu_{\rm cr}}^{\infty} d\nu e^{-\nu^2/2} = \frac{1}{2}\,{\rm erfc}\left(\frac{\nu_{\rm cr}}{\sqrt{2}}\right), \tag{A.152}$$

where we have used the definition of the complementary error function in Eq. (C.31). The joint probability is correspondingly given by two integrals over Eq. (12.111):

$$p(\delta_R^{(1)}(\boldsymbol{x}_1) > \delta_{\rm cr}, \delta_R^{(1)}(\boldsymbol{x}_2) > \delta_{\rm cr}) = \frac{1}{2\pi\sqrt{1-\hat{\xi}^2}} \int_{\nu_{\rm cr}}^{\infty} d\nu_1 \int_{\nu_{\rm cr}}^{\infty} d\nu_2 \exp\left[-\frac{1}{2}(\nu_1,\nu_2)^{\top} {\rm C}^{-1} (\nu_1,\nu_2)\right]. \tag{A.153}$$

Writing out the argument of the exponent, we have

$$-\frac{1}{2}(\nu_1,\nu_2)^{\top} {\rm C}^{-1} (\nu_1,\nu_2) = -\frac{\nu_1^2 + \nu_2^2 - 2\hat{\xi}\nu_1\nu_2}{2(1-\hat{\xi}^2)} = -\frac{1}{2}\left[w^2 + \nu_1^2\right] \tag{A.154}$$

where $w \equiv (\nu_2 - \hat{\xi}\nu_1)/\sqrt{1-\hat{\xi}^2}$. Switching integration variables from $\nu_2$ to $w$ yields

$$p(\delta_R^{(1)}(\boldsymbol{x}_1) > \delta_{\rm cr}, \delta_R^{(1)}(\boldsymbol{x}_2) > \delta_{\rm cr}) = \frac{1}{2\pi} \int_{\nu_{\rm cr}}^{\infty} d\nu_1 e^{-\nu_1^2/2} \int_{(\nu_{\rm cr}-\hat{\xi}\nu_1)/\sqrt{1-\hat{\xi}^2}}^{\infty} dw e^{-w^2/2}$$
$$= \frac{1}{\sqrt{2\pi}} \int_{\nu_{\rm cr}}^{\infty} d\nu_1 e^{-\nu_1^2/2} \frac{1}{2}\,{\rm erfc}\left[\frac{\nu_{\rm cr} - \hat{\xi}\nu_1}{\sqrt{2(1-\hat{\xi}^2)}}\right]. \tag{A.155}$$

Now, from Eq. (12.82), we simply have to divide by the one-point probability squared, yielding

$$1 + \xi_{\rm thr}(r) = \sqrt{\frac{2}{\pi}}\left[{\rm erfc}(\nu_{\rm cr}/\sqrt{2})\right]^{-2} \int_{\nu_{\rm cr}}^{\infty} d\nu_1 e^{-\nu_1^2/2} \frac{1}{2}\,{\rm erfc}\left[\frac{\nu_{\rm cr} - \hat{\xi}\nu_1}{\sqrt{2(1-\hat{\xi}^2)}}\right]. \tag{A.156}$$

This is the exact result for the correlation function of regions above threshold in a Gaussian density field.

**(c)** Eq. (A.156) still involves an integral that has to be done numerically. When looking at clustering at large separations $r$, we can do an expansion in $\hat{\xi}(r)$ whose magnitude is

much less than one. This yields derivatives of the error function erfc:

$$\begin{aligned}\mathrm{erfc}\left[\frac{\nu_{\rm cr}-\hat{\xi}\,\nu_1}{\sqrt{2(1-\hat{\xi}^2)}}\right] &= \mathrm{erfc}\left[\frac{\nu_{\rm cr}}{\sqrt{2}}\right] + \hat{\xi}\,\frac{\partial}{\partial\hat{\xi}}\,\mathrm{erfc}\left[\frac{\nu_{\rm cr}-\hat{\xi}\,\nu_1}{\sqrt{2(1-\hat{\xi}^2)}}\right]_0 + \cdots \\ &= \mathrm{erfc}\left[\frac{\nu_{\rm cr}}{\sqrt{2}}\right] + \sqrt{\frac{2}{\pi}}\,\nu_1 e^{\nu_{\rm cr}^2/2}\hat{\xi} + \cdots . \end{aligned} \tag{A.157}$$

Each higher derivative generates one more power of $\nu_1$ (more precisely, Hermite polynomials of $\nu_1$). Now we can perform the $\nu_1$ integral analytically. The zeroth-order term just cancels the 1 on the left-hand-side of Eq. (A.156). The first-order and second-order terms yield

$$\begin{aligned}\xi_{\rm thr}(r) &= (b_1^{\rm thr})^2\xi_R(r) + \frac{1}{2}(b_2^{\rm thr})^2[\xi_R(r)]^2 \quad \text{where} \\ b_1^{\rm thr} &= \sqrt{\frac{2}{\pi}}\,\frac{e^{-\nu_{\rm cr}^2/2}}{\mathrm{erfc}[\nu_{\rm cr}/\sqrt{2}]\sigma(R)} \overset{\nu_{\rm cr}\gg1}{\approx} \frac{\nu_{\rm cr}}{\sigma(R)} \\ b_2^{\rm thr} &= \sqrt{\frac{2}{\pi}}\,\frac{e^{-\nu_{\rm cr}^2/2}}{\mathrm{erfc}[\nu_{\rm cr}/\sqrt{2}]\sigma^2(R)}\,\nu_{\rm cr} \overset{\nu_{\rm cr}\gg1}{\approx} \frac{\nu_{\rm cr}^2}{\sigma^2(R)}. \end{aligned} \tag{A.158}$$

The factors of $\sigma(R)$ in the denominators here arise because we define the bias parameters as coefficients of $\xi_R(r)$ rather than $\hat{\xi}(r)$. The approximate expressions for rare high-density regions are also given. We see that rare excursions are highly biased, with the higher-order bias parameters growing more rapidly.

*Exercise 12.13*

**(a)** The convolution in real space becomes a multiplication in Fourier space. Define the power spectrum of halos of different masses as

$$\left\langle\delta_{\rm h}(\boldsymbol{k},M)\delta_{\rm h}(\boldsymbol{k}',M')\right\rangle = (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}')\,P_{\rm h}(k,M,M'). \tag{A.159}$$

Now we can perform the integrals over mass to obtain

$$P^{\rm HM}(k) = \frac{1}{\rho_{\rm m}^2}\int d\ln M\,\frac{dn}{d\ln M}M\int d\ln M'\,\frac{dn}{d\ln M'}M'\,y(k,M)y(k,M')P_{\rm h}(k,M,M'). \tag{A.160}$$

**(b)** The halo power spectrum Eq. (12.117) has two terms:

$$\begin{aligned}\left\langle\delta_{\rm h}(\boldsymbol{k},M)\delta_{\rm h}(\boldsymbol{k}',M')\right\rangle &= (2\pi)^3\delta_{\rm D}^{(3)}(\boldsymbol{k}+\boldsymbol{k}') \\ &\quad\times\left[b_1(M)b_1(M')P_{\rm L}(k) + \frac{1}{dn/d\ln M}\delta_{\rm D}^{(1)}(\ln M-\ln M')\right]. \end{aligned} \tag{A.161}$$

We can use this to break Eq. (A.160) into two terms:

$$P^{\rm HM}(k) = P_{\rm 2h}(k) + P_{\rm 1h}(k), \tag{A.162}$$

where

$$\begin{aligned} P_{\rm 2h}(k) &= [\mathcal{B}_1(k)]^2 P_{\rm L}(k), \\ \mathcal{B}_1(k) &= \frac{1}{\rho_{\rm m}} \int d\ln M \frac{dn}{d\ln M} M b_1(M) y(k, M), \\ P_{\rm 1h}(k) &= \frac{1}{\rho_{\rm m}^2} \int d\ln M \frac{dn}{d\ln M} M^2 [y(k, M)]^2. \end{aligned} \tag{A.163}$$

The prefactors of $1/\rho_{\rm m}$, $1/\rho_{\rm m}^2$ (all evaluated at $t_0$) are due to the fact that $P(k)$ is the power spectrum of the *fractional* matter density perturbation $\delta_{\rm m}$. Notice that a necessary condition is that $\lim_{k\to 0} \mathcal{B}(k) = 1$, which, using the profile normalization, requires

$$\int d\ln M \frac{dn}{d\ln M} M b_1(M) = 1. \tag{A.164}$$

It is thus important that this *bias consistency relation* is satisfied, which indeed is the case if the mass function is normalized such that all mass is in halos, and if $b_1$ is derived through the peak-background split, Eq. (12.80). The integral in Eq. (A.164) typically converges very slowly toward low masses. As argued in Appendix A of Schmidt (2016), however, the contribution from low-mass halos is trivial on the scales of interest. So one can simply cut off the mass integral and shift $\mathcal{B}_1(k)$ by an additive constant to enforce $\lim_{k\to 0} \mathcal{B}(k) = 1$. Similarly, the mass integral in $P_{\rm 1h}(k)$ can also be cut off.

**(c)** The Fourier transform of the halo profile is given by (see also Exercise 11.1)

$$y(k, M) = \frac{4\pi}{M} \int_0^{R_{200}} r^2 dr \frac{\sin kr}{kr} \rho_{\rm h}(r, M). \tag{A.165}$$

Notice that we truncate the profile at $R_{200}$ so that the total mass $M = M_{200}$ is finite. The result of the integral is

$$\begin{aligned} y\left(k = \frac{x}{r_s}, M\right) = \frac{1}{\mathcal{N}} \Big[ &\cos x\, [\mathrm{Ci}([c+1]x) - \mathrm{Ci}(x)] + \sin x\, [\mathrm{Si}([c+1]x) - \mathrm{Si}(x)] \\ &- \frac{\sin cx}{(c+1)x} \Big] \qquad \text{where} \\ \mathcal{N} = \frac{1}{c+1} &+ \log(c+1) - 1, \end{aligned} \tag{A.166}$$

and the scale radius $r_s$ is given in terms of the concentration $c$ by $r_s = R_{200}/c$. Ci and Si are the cosine and sine integrals, respectively, defined in Eqs. (C.32)–(C.33).

**(d)** The result is shown in Fig. 12.12.

One final comment: Eq. (A.163) shows that the 1-halo term asymptotes to a constant at large scales,

$$\lim_{k\to 0} P_{\rm 1h}(k) = \frac{1}{\rho_{\rm m}^2}\int d\ln M \frac{dn}{d\ln M} M^2. \tag{A.167}$$

This contribution is unphysical, since there cannot be any constant noise contribution to the matter power spectrum. The underlying physical error is the assumption of uncorrelated Poisson noise made in Eq. (12.117), which is inconsistent with the basic hypothesis of the halo model. Consider a box of cosmological volume. We know that large-scale density fluctuations are small, so the mean density within the box is close to the cosmic mean. Now we distribute this mass among halos. If, due to noise fluctuations, there are many massive halos, there have to be correspondingly fewer lower-mass halos, since there is a fixed amount of total matter. This constraint is violated if we assume independent noise in halos of all masses, as in Eq. (12.117). Fortunately, on the scales on which the halo model is applied, this is a numerically small contribution (Fig. 12.12), at least in the concordance cosmology.

# Chapter 13

*Exercise 13.1*

If the measurement of radiation happens at point $\boldsymbol{x}$ and time $t$ and at a frequency $\nu$, then we need to count the number of photons $dN$ with energy $E = p \in 2\pi[\nu, \nu + d\nu]$ arriving within a time interval $dt$ from a solid-angle element $d\Omega$ around a direction $\hat{\boldsymbol{n}} = -\hat{\boldsymbol{p}}$. This is

$$dN = 2f(\boldsymbol{x}, \boldsymbol{p}, t)\, dA_\perp dt \frac{d^3p}{(2\pi)^3} = 2f(\boldsymbol{x}, p = 2\pi\nu, \hat{\boldsymbol{p}}, t)\, dA_\perp\, dt\, \nu^2 d\nu\, d\Omega, \tag{A.168}$$

where the factor of 2 counts the two photon polarization states. We have used the fact that, since photons travel at speed $c = 1$, the detector collects photons from a volume element $d^3x = dA_\perp dt$ within a time interval $dt$. Further, since $p = 2\pi\nu$, the momentum-space volume element is $d^3p = p^2 dp d\Omega_p = (2\pi)^3\nu^2 d\nu d\Omega$. Weighting each photon by energy adds another factor of $E = 2\pi\nu$, so we obtain

$$I_\nu(\boldsymbol{x}, \boldsymbol{p}, t) = 4\pi\nu^3 f(\boldsymbol{x}, p = 2\pi\nu, \hat{\boldsymbol{p}}, t). \tag{A.169}$$

An equilibrium photon distribution has $f(p) = (\exp[p/k_{\rm B}T] - 1)^{-1}$, so the equilibrium intensity, more famously known as a black-body or Planck spectrum, is

$$I_\nu = \frac{4\pi\hbar\nu^3}{c^2}\left[\exp\left(\frac{2\pi\hbar\nu}{k_{\rm B}T}\right) - 1\right]^{-1}, \tag{A.170}$$

where we have restored the factors of $\hbar$ and $c$. This is Eq. (1.9).

*Exercise 13.4*

The solution to the Poisson equation (12.5) for an isolated mass is

$$\Phi(\boldsymbol{x}) = -Ga^2 \int \frac{d^3\tilde{x}}{|\boldsymbol{x}-\tilde{\boldsymbol{x}}|}\rho(\tilde{\boldsymbol{x}}). \tag{A.171}$$

Inserting this into Eq. (13.16) yields another integral, over $\chi'$. We will do both integrals in cylindrical coordinates, so that $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{R}}, \tilde{\chi})$ where $\tilde{\boldsymbol{R}}$ describes the transverse coordinates. Thus,

$$\phi_{\rm L}(\boldsymbol{\theta}; \chi_L) = -\frac{2G}{(1+z_L)^2}\frac{\chi-\chi_L}{\chi\chi_L}\int d^2\tilde{R}\int d\tilde{\chi}\,\rho(\tilde{\boldsymbol{R}},\tilde{\chi})\int_0^{\chi}\frac{d\chi'}{\sqrt{(\tilde{\boldsymbol{R}}-\chi_L\boldsymbol{\theta})^2+(\chi'-\tilde{\chi})^2}} \tag{A.172}$$

where we have set $\chi = \chi_L$ in the slowly varying factors in Eq. (13.16), and similarly evaluated $a^2$ at the lens redshift. This is accurate as long as the extent of the lens along the line of sight is small compared to the distance $\chi$ to the source galaxies, which certainly holds for individual galaxy clusters.

The integral over $d\chi'$ can be done analytically: it is equal to

$$2\,\ln\left|x+\sqrt{(\tilde{\boldsymbol{R}}-\chi_L\boldsymbol{\theta})^2+x^2}\right|\Bigg|_{x=0}^{\infty}$$

where we have set the upper limit to infinity because there is no contribution to the relevant part of the projected potential from large $x$. In fact, the only part which depends on $\boldsymbol{\theta}$ (and hence is relevant when derivatives of the lensing potential are taken) comes from the lower limit: $-2\ln|\tilde{\boldsymbol{R}}-\chi_L\boldsymbol{\theta}|$. We can pull out a factor $\chi_L$ from the logarithm for the same reason, since the additive term $\ln\chi_L$ has no dependence on $\boldsymbol{\theta}$. The integral over $\rho d\tilde{\chi}$ then becomes the surface density $\Sigma(\boldsymbol{\theta}')$, where $\boldsymbol{\theta}' = \tilde{\boldsymbol{R}}/\chi_L$. Using that $d^2\tilde{R} = \chi_L^2 d^2\theta'$, this leaves Eq. (13.67).

# Chapter 14

*Exercise 14.4*

First, double-decompose the general beam into spherical harmonics:

$$B(\hat{\boldsymbol{n}},\hat{\boldsymbol{n}}') = \sum_{lm,l'm'} B_{lm,l'm'}Y_{lm}(\hat{\boldsymbol{n}})Y^*_{l'm'}(\hat{\boldsymbol{n}}'), \tag{A.173}$$

where the use of the complex-conjugate spherical harmonic for $l'm'$ is by convention and for later convenience. Inserting this, and the decomposition of $\Theta$ into $a_{lm}$, into Eq. (14.36)

yields

$$\begin{aligned}\Delta(\hat{\boldsymbol{n}}) &= \int d\Omega' \sum_{l''m''} Y_{l''m''} a_{l''m''} \sum_{lm,l'm'} B_{lm,l'm'} Y_{lm}(\hat{\boldsymbol{n}}) Y^*_{l'm'}(\hat{\boldsymbol{n}}') \\ &= \sum_{lm} Y_{lm}(\hat{\boldsymbol{n}}) \sum_{l'm'} B_{lm,l'm'} a_{l'm'}. \end{aligned} \tag{A.174}$$

We can now read off the coefficient of $Y_{lm}(\hat{\boldsymbol{n}})$ as $a_{lm}^{\rm obs}$. This, with noise added, is Eq. (14.37).

If $B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') = B(\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{n}}')$, that is, the beam is only a function of the angle between the two directions, then we can do a Legendre decomposition instead, and use Eq. (C.12):

$$\begin{aligned} B(\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{n}}') &= \sum_l (2l+1) \tilde{B}_l \mathcal{P}_l(\hat{\boldsymbol{n}} \cdot \hat{\boldsymbol{n}}') \\ &= 4\pi \sum_{lm} \tilde{B}_l Y_{lm}(\hat{\boldsymbol{n}}) Y^*_{lm}(\hat{\boldsymbol{n}}'), \end{aligned} \tag{A.175}$$

so we have

$$B_{lm,l'm'} = B_l \delta_{ll'} \delta_{mm'} \tag{A.176}$$

where $B_l \equiv 4\pi \tilde{B}_l$. Inserting this into Eq. (14.37) immediately yields Eq. (14.38).

*Exercise 14.10*

**(a)** Let us expand the maximum-likelihood condition around $\bar{\lambda}$, the parameter value where it would peak if systematics were absent:

$$\begin{aligned} &\frac{d\ln\mathcal{L}}{d\lambda} = \frac{d\ln\mathcal{L}}{d\bar{\lambda}} + \frac{d^2\ln\mathcal{L}}{d\bar{\lambda}^2}(\lambda - \bar{\lambda}) = 0 \\ \Rightarrow \quad &\lambda = \bar{\lambda} + \mathcal{F}^{-1} \frac{d\ln\mathcal{L}}{d\bar{\lambda}}, \end{aligned} \tag{A.177}$$

since $d^2\ln\mathcal{L}/d\bar{\lambda}^2 = -\mathcal{F}$. Further, we have

$$\frac{d\ln\mathcal{L}}{d\bar{\lambda}} = \sum_l \frac{\partial C^{\rm theory}(l,\bar{\lambda})}{\partial\bar{\lambda}} \frac{\hat{C}(l) - C^{\rm theory}(l,\bar{\lambda})}{{\rm Var}\left[\hat{C}(l)\right]}. \tag{A.178}$$

**(b)** Now we can take the expectation value. $\mathcal{F}$ turns into $F$, while by assumption $\langle \hat{C}(l) - C^{\rm theory}(l,\bar{\lambda})\rangle = C^{\rm sys}(l)$: after averaging over many noise realizations, the mismatch between the observed $\hat{C}(l)$ and the theory $C^{\rm theory}(l,\bar{\lambda})$ evaluated for the true universe is nonzero and given by $C^{\rm sys}(l)$. We obtain

$$\lambda = \bar{\lambda} + F^{-1} \sum_l \frac{\partial C^{\rm theory}(l,\bar{\lambda})}{\partial\bar{\lambda}} \frac{C^{\rm sys}(l)}{{\rm Var}\left[\hat{C}(l)\right]}. \tag{A.179}$$

Notice that the sum over $l$ can be interpreted as a scalar product between $\partial C^{\rm theory}/\partial\bar{\lambda}$ and $C^{\rm sys}(l)$, weighted by the inverse variance. If there is little overlap between the systematic $C^{\rm sys}(l)$ and how the model $C^{\rm theory}(l)$ depends on $\bar{\lambda}$, then there will not be a significant shift in $\bar{\lambda}$ (we say that $C^{\rm sys}(l)$ is "almost orthogonal" to $\bar{\lambda}$). Consider for example a systematic that is essentially constant in $l$, and a parameter (e.g., $\Omega_{\rm b}h^2$) that leads to an oscillatory change in $C^{\rm theory}(l)$. Then, we expect this systematic to lead to a negligible parameter shift. Eq. (A.179) captures this effect rigorously.

**(c)** The generalization to multiple parameters is straightforward. Eq. (A.177) becomes a vectorial relation. If we define

$$B_\alpha = \sum_l \frac{\partial C^{\rm theory}(l,\{\bar{\lambda}_\gamma\})}{\partial\bar{\lambda}_\alpha}\frac{C^{\rm sys}(l)}{{\rm Var}\left[\hat{C}(l)\right]}, \tag{A.180}$$

we obtain

$$\lambda_\alpha = \bar{\lambda}_\alpha + \left(F^{-1}\right)_{\alpha\beta} B_\beta. \tag{A.181}$$

The interpretation is the same as in the one-parameter case, except that now a shift in the parameter $\lambda_\alpha$ can occur even if $B_\alpha$ is small. This happens if $B_\beta$ is significant for other parameters $\lambda_\beta$ that are partially degenerate with $\lambda_\alpha$, so that $(F^{-1})_{\alpha\beta} \neq 0$. We see that the parameter degeneracies encoded by the Fisher matrix are also very important in this case.

# B

# Numbers

Numbers in parentheses denote one standard deviation uncertainties in last digits (e.g., for Rydberg's constant, $\epsilon_0 = (13.60569172 \pm 5.3 \times 10^{-7})$ eV). The majority of the numbers quoted here come from the Particle Data Group (Tanabashi et al., 2018), while the fiducial cosmology is taken from Planck collaboration (2018b).

## B.1 Physical constants

| | | | |
|---|---|---|---|
| Speed of light | $c$ | = | $2.99792458 \times 10^{10}$ cm s$^{-1}$ |
| Reduced Planck's constant | $\hbar$ | = | $6.58211889(26) \times 10^{-16}$ eV s |
| | | = | $1.973269602(77) \times 10^{-5}$ eV cm$/c$ |
| Newton's constant | $G$ | = | $6.673(10) \times 10^{-8}$ cm$^3$ g$^{-1}$ s$^{-2}$ |
| | | = | $\hbar c/m_{\rm Pl}^2$ |
| Planck mass | $m_{\rm Pl}$ | = | $\sqrt{\hbar c/G}$ |
| | | = | $1.221 \times 10^{19}$ GeV$/c^2$ |
| | | = | $1.094 \times 10^{-38} M_\odot$ |
| Boltzmann constant | $k_{\rm B}$ | = | $8.617342(15) \times 10^{-5}$ eV K$^{-1}$ |
| Fine structure constant | $\alpha$ | = | $1/137.03599976(50)$ |
| Electron mass | $m_e$ | = | $0.510998902(21)$ MeV$/c^2$ |
| Ground-state energy of hydrogen | $\epsilon_0$ | = | $m_e c^2 \alpha^2/2$ |
| (Rydberg's constant) | | = | 13.60569172(53) eV |
| Thomson cross-section | $\sigma_{\rm T}$ | = | $8\pi\alpha^2\hbar^2/3m_e^2c^2$ |
| | | = | $0.665245854(15) \times 10^{-24}$ cm$^2$ |
| Neutron mass | $m_n$ | = | $939.565330(38)$ MeV$/c^2$ |
| Proton mass | $m_p$ | = | $1.67262158(13) \times 10^{-24}$ g |
| | | = | $938.271998(38)$ MeV$/c^2$ |
| Neutron–proton mass difference | $\mathcal{Q}$ | = | 1.2933 MeV$/c^2$ |
| Neutron lifetime | $\tau_n$ | = | 885.7(8) s |
| Fermi constant | $G_{\rm F}$ | = | $1.16639(1) \times 10^{-5}$ GeV$^{-2}(\hbar c)^3$ |

## B.2 Astrophysical constants

| | | | |
|---|---|---|---|
| Cosmic microwave background | $\rho_\gamma$ | = | $\pi^2 k_{\rm B}^4 T^4/15(\hbar c)^3$ |
| energy density | | = | $2.474 \times 10^{-5} h^{-2} (T/T_0)^4 \rho_{\rm cr}$ |
| Critical density | $\rho_{\rm cr}$ | = | $1.879\, h^2 \times 10^{-29}$ g cm$^{-3}$ |
| | | = | $2.775\, h^2 \times 10^{11} M_\odot$ Mpc$^{-3}$ |
| | | = | $8.098\, h^2 \times 10^{-11}$ eV$^4/(\hbar c)^3$ |

| | | | |
|---|---|---|---|
| Neutrino density parameter today | $\Omega_\nu h^2$ | $=$ | $\sum m_\nu/94\ \mathrm{eV}$ |
| Scale factor at equality | $a_{\mathrm{eq}}$ | $=$ | $4.15\times10^{-5}\,(\Omega_{\mathrm{m}}h^2)^{-1}$ |
| Inverse comoving horizon ($\eta^{-1}$) at equality | $k_{\mathrm{eq}}$ | $=$ | $0.073\,\Omega_{\mathrm{m}}h^2\,\mathrm{Mpc}^{-1}$ |
| Hubble constant | $H_0$ | $=$ | $100h\ \mathrm{km\ s^{-1}\,Mpc^{-1}}$ |
| | | $=$ | $2.133h\times10^{-42}\ \mathrm{GeV}/\hbar$ |
| | | $=$ | $1.023h\times10^{-10}\ \mathrm{yr}^{-1}$ |
| Solar mass | $M_\odot$ | $=$ | $1.989\times10^{33}\ \mathrm{g}$ |
| | | $=$ | $1.116\times10^{57}\ \mathrm{GeV}/c^2$ |
| Parsec | pc | $=$ | $3.0856\times10^{18}\ \mathrm{cm}$ |
| Cosmic microwave background temperature today | $T_0$ | $=$ | $2.726(1)\ \mathrm{K}$ |
| | | $=$ | $2.349\times10^{-4}\ \mathrm{eV}/k_{\mathrm{B}}$ |

## B.3 Fiducial cosmology

**Table B.1** Fiducial cosmology used throughout this book. It is the `base_plikHM_TTTEEE_lowl_lowE_lensing_post_BAO` best-fit Euclidean ΛCDM cosmology from Planck Collaboration (2018b) (see parameters document and wiki). The six parameters in the upper part are the primary parameters of the concordance cosmology, while those below are derived parameters. The last column lists the 95% confidence-level parameter limits.

| Parameter | Symbol | Best fit | 95% C.L. limits |
|---|---|---|---|
| Baryon density parameter | $\Omega_{\mathrm{b}}h^2$ | 0.022447 | $\pm0.00027$ |
| CDM density parameter | $\Omega_{\mathrm{c}}h^2$ | 0.11928 | $\pm0.0018$ |
| Optical depth due to reionization | $\tau_{\mathrm{rei}}$ | 0.0568 | $\pm0.014$ |
| Hubble parameter | $h$ | 0.6770 | $\pm0.0081$ |
| Scalar spectral index | $n_s$ | 0.9682 | $+0.0076/-0.0073$ |
| Scalar power spectrum amplitude | $\ln(10^{10}\mathcal{A}_s)$ | 3.0480 | $\pm0.028$ |
| Cosmological constant parameter | $\Omega_\Lambda$ | 0.6894 | $\pm0.011$ |
| Matter density parameter | $\Omega_{\mathrm{m}}$ | 0.3106 | $\pm0.011$ |
| Matter power spectrum normalization at $t_0$ (Fig. 12.1) | $\sigma_8$ | 0.8110 | $\pm0.012$ |
| Age of the universe [Gyr] | $t_0$ | 13.784 | $+0.040/-0.037$ |

# C Special functions

This appendix provides a very brief summary of special functions that are relevant in the cosmological context. For a more complete treatment, see, e.g., the *Handbook of Mathematical Functions* (Abramowitz and Stegun).

## C.1 Legendre polynomials

The Legendre polynomial $\mathcal{P}_l(\mu)$ is an $l$th order polynomial in $\mu$. For $-1 \le \mu \le 1$, $\mathcal{P}_l$ has $l$ zeroes in this interval. The first few polynomials are

$$\begin{aligned} \mathcal{P}_0(\mu) &= 1, \\ \mathcal{P}_1(\mu) &= \mu, \\ \mathcal{P}_2(\mu) &= \frac{3\mu^2 - 1}{2}. \end{aligned} \tag{C.1}$$

The property observed for these lowest $l$, that $\mathcal{P}_l$ is an even function of $\mu$ for $l$ even and an odd function for $l$ odd, holds for all $l$. The Legendre polynomials are orthogonal on the interval $[-1, 1]$, so that

$$\int_{-1}^{1} d\mu \mathcal{P}_l(\mu)\mathcal{P}_{l'}(\mu) = \delta_{ll'} \frac{2}{2l+1}. \tag{C.2}$$

In fact, they form a complete basis on this interval. To generate the higher-order Legendre polynomials, one can use the recurrence relation

$$(l+1)\mathcal{P}_{l+1}(\mu) = (2l+1)\mu\mathcal{P}_l(\mu) - l\mathcal{P}_{l-1}(\mu). \tag{C.3}$$

This relation is useful for expanding the Boltzmann equations in terms of moments.

## C.2 Spherical harmonics

Spherical harmonics are eigenfunctions of the angular part of the Laplacian,

$$\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right] Y_{lm}(\theta,\phi) = -l(l+1)Y_{lm}(\theta,\phi). \tag{C.4}$$

The CMB temperature is defined on the sphere, i.e., is a function of $\theta$, $\phi$, so it is naturally expanded in $Y_{lm}$ (Eq. (9.63)). This decomposition is the analogue of a two-dimensional

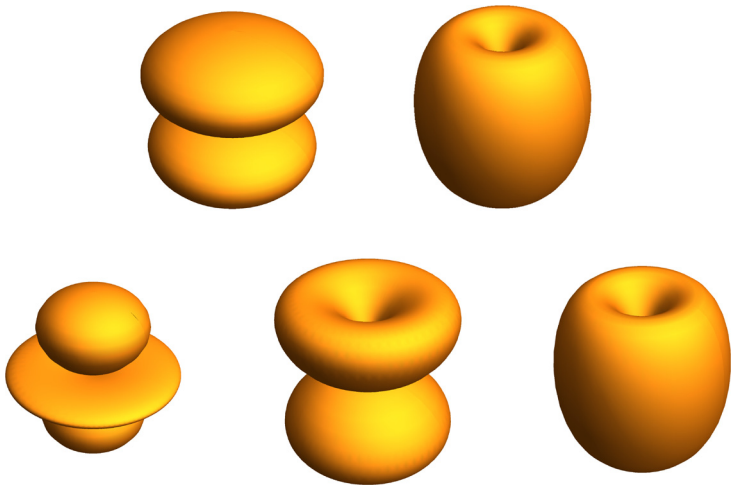


**FIGURE C.1** 3D contour plots of the absolute value of spherical harmonics for $l = 1$ (top row) and $l = 2$ (bottom row). The $z$-axis points upwards. *Top row, from left:* $|Y_{10}|$, $|Y_{11}|$. *Bottom row, from left:* $|Y_{20}|$, $|Y_{21}|$, $|Y_{22}|$.

Fourier decomposition in flat space. The lowest spherical harmonics are

$$Y_{00}(\theta, \phi) = \frac{1}{\sqrt{4\pi}}, \tag{C.5}$$

$$Y_{10}(\theta, \phi) = i\sqrt{\frac{3}{4\pi}}\cos(\theta), \tag{C.6}$$

$$Y_{1,\pm1}(\theta, \phi\rangle = \mp i\sqrt{\frac{3}{8\pi}}\sin(\theta)e^{\pm i\phi}, \tag{C.7}$$

$$Y_{20}(\theta, \phi) = \sqrt{\frac{5}{16\pi}}(1 - 3\cos^2\theta), \tag{C.8}$$

$$Y_{2,\pm1}(\theta, \phi) = \pm i\sqrt{\frac{15}{8\pi}}\cos\theta\sin\theta e^{\pm i\phi}, \tag{C.9}$$

$$Y_{2,\pm2}(\theta, \phi) = -\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{\pm 2i\phi}. \tag{C.10}$$

Contour plots of the absolute values of some of these spherical harmonics are shown in Fig. C.1.

The spherical harmonics form a complete basis on the sphere and are orthogonal, with normalization

$$\int d\Omega Y^*_{lm}(\hat{\boldsymbol{n}})Y_{l'm'}(\hat{\boldsymbol{n}}) = \delta_{ll'}\delta_{mm'}. \tag{C.11}$$

Another useful expression is the Legendre polynomial in terms of a sum of products of the spherical harmonics:

$$\mathcal{P}_l(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{n}}') = \frac{4\pi}{2l+1}\sum_{m=-l}^{l} Y_{lm}(\hat{\boldsymbol{n}})Y^*_{lm}(\hat{\boldsymbol{n}}'). \tag{C.12}$$

## C.3 Spherical Bessel functions

Spherical Bessel functions are crucial in the study of the CMB and large-scale structure in particular because they appear when projecting inhomogeneities onto the sky. They satisfy the differential equation

$$\frac{d^2 j_l}{dx^2} + \frac{2}{x}\frac{dj_l}{dx} + \left[1 - \frac{l(l+1)}{x^2}\right] j_l = 0. \tag{C.13}$$

The lowest several are

$$j_0(x) = \frac{\sin(x)}{x}; \quad j_1(x) = \frac{\sin x - x\cos x}{x^2}. \tag{C.14}$$

The key integral relating Legendre polynomials to spherical Bessel functions is

$$\frac{1}{2}\int_{-1}^{1} d\mu\, \mathcal{P}_l(\mu)e^{iz\mu} = \frac{j_l(z)}{(-i)^l}. \tag{C.15}$$

Using the completeness of the Legendre polynomials, we can invert this relation to obtain a useful expansion of a plane-wave perturbation:

$$e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = \sum_{l=0}^{\infty} i^l (2l+1)\, j_l(kx)\mathcal{P}_l(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{x}}). \tag{C.16}$$

Combining with Eq. (C.12) yields

$$e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = 4\pi\sum_{l=0}^{\infty} i^l\, j_l(kx)\sum_{m=-l}^{l} Y_{lm}(\hat{\boldsymbol{k}})Y^*_{lm}(\hat{\boldsymbol{x}}). \tag{C.17}$$

An important integral, useful for computing the Sachs–Wolfe effect, is

$$\int_0^\infty dx\, x^{n-2}[j_l(x)]^2 = 2^{n-4}\pi\frac{\Gamma(l+\frac{n}{2}-\frac{1}{2})\Gamma(3-n)}{\Gamma(l+\frac{5}{2}-\frac{n}{2})\Gamma^2(2-\frac{n}{2})}, \tag{C.18}$$

where the $\Gamma$ function is defined in App. C.5.

Another important relation to eliminate derivatives of Bessel functions is

$$\frac{dj_l}{dx} = j_{l-1} - \frac{l+1}{x} j_l. \tag{C.19}$$

Finally, the following recurrence relation is useful also in numerical implementations:

$$j_{l+1}(x) = \frac{2l+1}{x} j_l(x) - j_{l-1}(x). \tag{C.20}$$

## C.4 Fourier transforms

Our Fourier convention is

$$\begin{aligned} f(\boldsymbol{x}) &= \int \frac{d^3k}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \tilde{f}(\boldsymbol{k}), \\ \tilde{f}(\boldsymbol{k}) &= \int d^3x\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} f(\boldsymbol{x}). \end{aligned} \tag{C.21}$$

The power spectrum is then the Fourier transform of the correlation function, with

$$\langle \tilde{\delta}(\boldsymbol{k}) \tilde{\delta}(\boldsymbol{k}') \rangle = (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k} + \boldsymbol{k}') P(k). \tag{C.22}$$

Since $\tilde{\delta}$ is the Fourier transform of a real field, we have $\tilde{\delta}(-\boldsymbol{k}) = \tilde{\delta}^*(\boldsymbol{k})$, so

$$\langle \tilde{\delta}(\boldsymbol{k}) \tilde{\delta}^*(\boldsymbol{k}') \rangle = (2\pi)^3 \delta_{\rm D}^{(3)}(\boldsymbol{k} - \boldsymbol{k}') P(k). \tag{C.23}$$

Notice that we mostly drop the tilde over Fourier-space variables in the text, since no confusion can arise.

## C.5 Miscellaneous

We just need a couple of relations involving ordinary Bessel functions,

$$J_n(x) = \frac{i^{-n}}{\pi} \int_0^\pi d\theta\, e^{ix\cos\theta} \cos(n\theta) \tag{C.24}$$

and

$$\frac{d}{dx}[x J_1(x)] = x J_0(x). \tag{C.25}$$

The $\Gamma$ function is related to factorials, in that it satisfies, for integer arguments,

$$\Gamma(n+1) = n!\,. \tag{C.26}$$

More generally, for any real or complex number $x$,

$$\Gamma(x+1) = x\Gamma(x). \tag{C.27}$$

The Sachs–Wolfe integral (Eq. (C.18)) for a scale-invariant spectrum ($n_s = 1$) involves

$$\Gamma(3/2) = \frac{\sqrt{\pi}}{2}. \tag{C.28}$$

The Riemann zeta function is useful for evaluating integrals in statistical mechanics. In particular,

$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty dx \frac{x^{s-1}}{e^x - 1} = \frac{1}{(1 - 2^{1-s})\Gamma(s)} \int_0^\infty dx \frac{x^{s-1}}{e^x + 1}. \tag{C.29}$$

The cases we encounter in this book are

$$\zeta(2) = \frac{\pi^2}{6}; \quad \zeta(3) = 1.202; \quad \zeta(4) = \frac{\pi^4}{90}. \tag{C.30}$$

When dealing with Gaussian random fields, one often encounters the error function erf, the incomplete integral over the Gaussian distribution, and its complement, erfc:

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty du\, e^{-u^2}. \tag{C.31}$$

The cosine and sine integrals appear when computing the Fourier transform of the NFW halo profile:

$$\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos z}{z} dz, \tag{C.32}$$

$$\mathrm{Si}(x) = \int_0^x \frac{\sin z}{z} dz. \tag{C.33}$$

# D

# Symbols

## D.1 Mathematical and geometrical definitions

| Symbol | Explanation |
|---|---|
| $\dot{f}(\boldsymbol{x},t) \equiv \partial f(\boldsymbol{x},t)/\partial t$ | Partial derivative with respect to time |
| $f'(\boldsymbol{x},\eta) \equiv \partial f(\boldsymbol{x},\eta)/\partial \eta$ | Derivative with respect to conformal time |
| $\phi_{,\alpha} \equiv \partial\phi(x)/\partial x^{\alpha}$ | Partial derivative with respect to coordinate $x^{\alpha}$ |
| ${\delta^{\nu}}_{\alpha}$, $\delta_{ij}$ | Kronecker symbol |
| $\delta_{\mathrm{D}}^{(n)}(\boldsymbol{k}-\boldsymbol{k})$ | Dirac-delta distribution in $n$ dimensions |
| $\hat{\boldsymbol{e}}_{x,y,z}$ | Unit vector in direction of three spatial Cartesian axes |
| $\hat{\boldsymbol{n}}$ | 3D unit vector (full-sky position) |
| $\boldsymbol{\theta}$ | 2D Euclidean vector (flat-sky position) |
| $d\Omega$ | Solid angle integration measure |

Throughout, spatial indices $ijk\ldots$ are raised and lowered with $\delta_{ij}$.

## D.2 Frequently used relations

Frequently used time integration measures are

$$d\eta = \frac{dt}{a(t)} = \frac{da}{a^2 H(a)} = \frac{d\ln a}{aH(a)}. \tag{D.1}$$

For light rays, we further have

$$d\chi = -d\eta = \frac{dz}{H(z)}. \tag{D.2}$$

Our convention for the perturbed FLRW metric is (Eq. (3.49))

$$\begin{aligned} g_{00}(\boldsymbol{x},t) &= -1 - 2\Psi(\boldsymbol{x},t), \\ g_{0i}(\boldsymbol{x},t) &= 0, \\ g_{ij}(\boldsymbol{x},t) &= a^2(t)\delta_{ij}\left[1 + 2\Phi(\boldsymbol{x},t)\right]. \end{aligned} \tag{D.3}$$

## D.3 Symbol definitions

| Symbol | Explanation | Defining equation |
|---|---|---|
| $a(t)$ | Scale factor | Eq. (2.12) |
| $z$ | Redshift | Eq. (1.1) |
| $t_0$ | Age of the universe / today's epoch | |
| $T_0$ | CMB temperature today | |
| $H(t)$, $H_0$ | Hubble rate, Hubble's constant | Eq. (1.2), $H_0 \equiv H(t_0)$ |
| $\rho(t)$ | Total background energy density | Eq. (2.44) |
| $\rho_{\rm cr}$ | Critical density today | Eq. (1.4) |
| $I_\nu$ | Specific intensity of radiation | Eq. (1.9) |
| $\overline{n}_{\rm g}$ | Mean galaxy number density | Above Eq. (1.10) |
| $\delta_{\rm g}$ | Fractional galaxy density perturbation | Above Eq. (1.10) |
| $g_{\mu\nu}$ | Spacetime metric | Eq. (2.4) |
| $\eta_{\mu\nu}$ | Minkowski metric | Eq. (2.11) |
| $\Gamma^\mu{}_{\alpha\beta}$ | Christoffel symbol | Eq. (2.21) |
| $P^\alpha$ | Comoving four-momentum | Eq. (2.26) |
| $p^i$ | Physical three-momentum | Eqs. (2.32), (3.28) |
| $\hat{\boldsymbol{p}}$ | Unit momentum vector | Eq. (3.32) |
| $\eta$, $\eta_0$ | Conformal time, value today | Eq. (2.35), $\eta_0 \equiv \eta(t_0)$ |
| $\chi(z)$ | Comoving distance out to redshift $z$ | Eq. (2.34) |
| $d_A(z)$ | Angular diameter distance | Eqs. (2.37), (2.39) |
| $d_L(z)$ | Luminosity distance | $d_L = d_A/a^2$ |
| $T^\mu{}_\nu$ | Energy-momentum tensor | Eq. (2.44) |
| $\mathcal{P}$ | Homogeneous pressure | Eq. (2.44) |
| $w$ | Equation of state | Eq. (2.60) |
| $\rho_{\rm m}(t)$, $\rho_{\rm r}(t)$ | Total matter and radiation (photon+neutrino) densities | Sect. 2.3 |
| $E_s(p)$ | Energy–momentum relation for species $s$ | $E_s = \sqrt{p^2 + m_s^2}$ |
| $g_s$ | Degeneracy factor for species $s$ | Below Eq. (2.62) |
| $f_{\rm BE}(E)$ | Bose–Einstein distribution | Eq. (2.65) |
| $f_{\rm FD}(E)$ | Fermi–Dirac distribution | Eq. (2.66) |
| $s$ | Entropy density *(only in Ch. 2 & Ch. 4)* | Eq. (2.70) |
| $\mu$ | *Chs. 2–4:* Chemical potential | |
| | *Chs. 5–12:* Cosine of angle between wavevector and photon momentum | Eq. (5.31) |
| | *Ch. 13:* magnification | Eq. (13.35) |
| $\Omega_s$ | Density parameter for constituent $s$ at $t_0$ | Eq. (2.71) |
| $a_{\rm eq}$ | Scale factor at matter–radiation equality | Eq. (2.86) |

| Symbol | Explanation | Defining equation |
|---|---|---|
| $G_{\mu\nu}$ | Einstein tensor | Eq. (3.2) |
| $R_{\mu\nu}$ | Ricci tensor | Eq. (3.3) |
| $R$ | *Ch. 3 & Ch. 6:* Ricci scalar | $R \equiv g^{\mu\nu} R_{\mu\nu}$ |
| $R(\eta)$ | *Ch. 5, Chs. 8–9:* Baryon/photon energy ratio | Eq. (5.74) |
| $\Psi$ | Perturbation to $g_{00}$ | Eq. (3.49) |
| $\Phi$ | Perturbation to $g_{ij}$ | Eq. (3.49) |
| $df(\boldsymbol{x}, \boldsymbol{p}, t)/dt$ | Total time derivative (phase space) | Eq. (3.17) |
| $C[f]$ | Collision term | Eqs. (3.19), (3.48) |
| $\mathcal{M}$ | Scattering amplitude | Eq. (3.46) |
| $n_s^{(0)}$ | Equilibrium number density of species $s$ | Eq. (4.5) |
| $\langle \sigma v \rangle$ | Thermally averaged cross section | Eq. (4.7) |
| $\eta_{\rm b}$ | Ratio of baryon to photon number density | Eq. (4.10) |
| $Y_P$ | Primordial $^4$He mass fraction | Eq. (4.30) |
| $\Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, t)$ | Temperature perturbation to photon distribution function | Eq. (5.2) |
| $\Theta_0(\boldsymbol{x}, t)$ | Monopole temperature perturbation | Eq. (5.20) |
| $\Theta_l(k, \eta)$ | Multipole moment of Fourier-space temperature perturbation | Eq. (5.66) |
| $u_{\rm c}$, $u_{\rm b}$ | Bulk velocity of CDM, baryons | Eqs. (5.39), (5.54) |
| $\delta_{\rm c}$, $\delta_{\rm b}$ | Fractional density pert. of CDM, baryons | Eqs. (5.44), (5.53) |
| $\mathcal{N}(\boldsymbol{x}, \boldsymbol{p}, t)$ | Perturbation to neutrino distribution function | Eq. (5.62) |
| $\mathcal{N}_l(k, \eta)$ | Multipole moment of $\mathcal{N}$ for massless neutrinos | Eq. (5.66) |
| $h_{ij}^{\rm TT}$ | Tensor metric perturbation (transverse-traceless) | Eqs. (6.6), (6.49) |
| $\Phi_A$, $\Phi_H$ | Bardeen's gauge-invariant perturbations | Eq. (6.19) |
| $\delta_{\rm m}$, $u_{\rm m}$ | Total matter density perturbation and velocity | Eq. (6.79) |
| $\Theta_{\rm r,0}$, $\Theta_{\rm r,1}$ | Total radiation monopole and dipole | Eq. (6.79) |
| $\eta_*$ | Conformal time (comoving horizon) at last scattering | |
| $H_{\rm inf}$ | Hubble parameter during inflation | Eq. (7.4) |
| $\epsilon_{\rm sr}$ | First slow-roll parameter | Eq. (7.17) |
| $\delta_{\rm sr}$ | Second slow-roll parameter | Eq. (7.18) |
| $\mathcal{R}$ | Curvature perturbation in comoving gauge | Eq. (7.57) |
| $\mathcal{A}_s$, $n_s$ | Primordial power spectrum normalization and index | Eq. (7.99) |
| $k_{\rm p}$ | Pivot scale | $k_{\rm p} = 0.05\ {\rm Mpc}^{-1}$ |
| $r$, $n_{\rm T}$ | Tensor-to-scalar ratio and tensor index | Eqs. (7.103), (7.102) |

| Symbol | Explanation | Defining equation |
|---|---|---|
| $T(k)$ | Matter transfer function | Eq. (8.2) |
| $D_+(a)$ | Linear growth factor | Eq. (8.3) |
| $P_{\rm L}(k,a)$ | Linear matter power spectrum | Eq. (8.8) |
| $\Delta^2_{\rm L}(k,a)$ | Dimensionless linear power spectrum | Eq. (8.9) |
| $k_{\rm NL}(a)$ | Nonlinear scale | $\Delta^2_{\rm L}(k_{\rm NL},a)=1$ |
| $\Omega_{\rm m}(a)$ | Time-dependent density parameter (only used in Sect. 8.5 and Ch. 12) | Below Eq. (8.78) |
| $y$ | *Ch. 8:* Scale factor in units of $a_{\rm eq}$ | Eq. (8.20) |
| | *Ch. 11:* SZ distortion parameter | Eq. (11.59) |
| $a_{lm}$ | Multipole moments of CMB temperature | Eq. (9.63) |
| $C(l)$ | Angular CMB power spectrum | Eq. (9.66) |
| $\mathcal{D}_l$ | Scaled CMB power spectrum | $\equiv l(l+1)C(l)T_0^2/2\pi$ |
| $\tau(\eta)$ | Compton-scattering optical depth | Eq. (5.33) |
| $\tau_{\rm rei}$ | Optical depth due to reionization | Sect. 9.7.2 |
| $g(\eta)$ | Visibility function | Eq. (9.56) |
| $I$, $Q$, $U$ | Stokes parameters | Eq. (10.2) |
| $E(\boldsymbol{l})$, $B(\boldsymbol{l})$ | $E$-mode and $B$-mode for angular wavenumber $\boldsymbol{l}$ | Eqs. (10.6), (10.9) |
| $C_{EE}(l)$, $C_{BB}(l)$ | *Ch. 10:* CMB polarization power spectra | Sect. 10.5 |
| | *Ch. 13:* Lensing shear power spectra | Sect. 13.5.1 |
| $C_{TE}(l)$ | Temperature-polarization cross power spectrum | Eq. (10.46) |
| $P_{\rm g,obs}(\boldsymbol{k},z)$ | Observed three-dimensional galaxy power spectrum | Eq. (11.37) |
| $C_{\rm g}(l)$ | Angular galaxy power spectrum | Eq. (11.43) |
| $P(k)$ | Nonlinear matter power spectrum | Eq. (C.22) for $\delta_{\rm m}$ |
| $\theta_{\rm m}$ | Matter velocity divergence | $\theta_{\rm m}=\partial_i u^i_{\rm m}$ |
| $R_\Delta$, $M_\Delta$ | Spherical-overdensity halo radius, mass | Eq. (12.61) |
| $R_L(M)$ | Halo Lagrangian radius | Eq. (12.64) |
| $dn/d\ln M$ | Halo mass function | Eq. (12.73) |
| $\phi_{\rm L}$ | Lensing potential | Eq. (13.16) |
| $\kappa$, $\gamma_1$, $\gamma_2$ | Lensing convergence and shear | Eq. (13.28) |
| $\gamma_t$, $\gamma_\times$ | Tangential and cross components of shear | Above Eq. (13.57) |
| $C_{{\rm g}E}(l)$ | Galaxy-shear cross power spectrum | Eq. (13.61) |
| $\mathcal{L}(\{d_i\}_{i=1}^m\|w,\sigma_w)$ | Likelihood function of data $\{d_i\}$ given parameters $w$, $\sigma_w$ | E.g., Eq. (14.2) |
| $P(w,\sigma_w\|\{d_i\}_{i=1}^m)$ | Posterior for parameters $w$, $\sigma_w$ given data $\{d_i\}$ | Eq. (14.5) |
| $F_{\alpha\beta}$ | Fisher information matrix | Eq. (14.69) |

# Bibliography


Abazajian, K.N., et al., 2016. CMB-S4 Science Book, first edition.

Abbott, T.M.C., et al., 2018. Dark energy survey year 1 results: cosmological constraints from galaxy clustering and weak lensing. Physical Review D 98 (4), 043526.

Ackermann, M., et al., 2014. Dark matter constraints from observations of 25 Milky Way satellite galaxies with the Fermi large area telescope. Physical Review D 89, 042001.

Ade, P.A.R., et al., 2018. BICEP2 / Keck array X: constraints on primordial gravitational waves using Planck, WMAP, and new BICEP2/Keck observations through the 2015 season. Physical Review Letters 121, 221301.

Aghanim, N., et al., 2014. Planck 2013 results. IV. Low frequency instrument beams and window functions. Astronomy & Astrophysics 571, A4.

Albrecht, A., Steinhardt, P.J., 1982. Cosmology for grand unified theories with radiatively induced symmetry breaking. Physical Review Letters 48, 1220–1223.

Alcock, C., Paczyński, B., 1979. An evolution free test for non-zero cosmological constant. Nature (London) 281, 358.

Anderson, L., et al., 2012. The clustering of galaxies in the SDSS-III baryon oscillation spectroscopic survey: baryon acoustic oscillations in the data release 9 spectroscopic galaxy sample. Monthly Notices of the Royal Astronomical Society 427, 3435–3467.

Audi, G., Wapstra, A.H., Thibault, C., 2003. The Ame2003 atomic mass evaluation: (II). Tables, graphs and references. Nuclear Physics. A 729 (1), 337–676. The 2003 NUBASE and Atomic Mass Evaluations.

Bahcall, J.N., 1989. Neutrino Astrophysics.

Bardeen, J.M., 1980. Gauge-invariant cosmological perturbations. Physical Review D 22, 1882–1905.

Bartelmann, M., Schneider, P., 2001. Weak gravitational lensing. Physics Reports 340, 291–472.

Baumann, D., Nicolis, A., Senatore, L., Zaldarriaga, M., 2012. Cosmological non-linearities as an effective fluid. J. Cosmol. Astropart. Phys. 7, 051. https://doi.org/10.1088/1475-7516/2012/07/051. arXiv:1004.2488.

Bennett, C.L., et al., 1996. Four-year COBE DMR cosmic microwave background observations: maps and basic results. The Astrophysical Journal Letters 464, L1.

Bernardeau, F., et al., 2002. Large scale structure of the universe and cosmological perturbation theory. Physics Reports 367, 1–248.

Bernstein, J., 2004. Kinetic Theory in the Expanding Universe.

Bernstein, J., Brown, L.S., Feinberg, G., 1989. Cosmological helium production simplified. Reviews of Modern Physics 61, 25.

Bertschinger, E., Jain, B., 1994. Gravitational instability of cold matter. The Astrophysical Journal 431, 486–494.

Beutler, F., et al., 2017. The clustering of galaxies in the completed SDSS-III baryon oscillation spectroscopic survey: baryon acoustic oscillations in the Fourier space. Monthly Notices of the Royal Astronomical Society 464, 3409–3430.

Birrell, N.D., Davies, P.C.W., 1984. Quantum Fields in Curved Space. Cambridge Monographs on Mathematical Physics. Cambridge Univ. Press, Cambridge, UK.

Blas, D., Lesgourgues, J., Tram, T., 2011. The cosmic linear anisotropy solving system (CLASS). Part II: Approximation schemes. Journal of Cosmology and Astroparticle Physics 7, 034.

Bleem, L.E., et al., 2015. Galaxy clusters discovered via the Sunyaev-Zel'dovich effect in the 2500-square-degree SPT-SZ survey. The Astrophysical Journal. Supplement Series 216, 27.

Bond, J.R., Efstathiou, G., 1984. Cosmic background radiation anisotropies in universes dominated by nonbaryonic dark matter. The Astrophysical Journal Letters 285, L45–L48.

Bond, J.R., Efstathiou, G., 1987. The statistics of cosmic background radiation fluctuations. Monthly Notices of the Royal Astronomical Society 226, 655–687.

Bond, J.R., Efstathiou, G., Silk, J., 1980. Massive neutrinos and the large-scale structure of the universe. Physical Review Letters 45, 1980–1984.

Bond, J.R., et al., 1991. Excursion set mass functions for hierarchical Gaussian fluctuations. The Astrophysical Journal 379, 440–460.

Bouwens, R.J., et al., 2015. Reionization after Planck: the derived growth of the cosmic ionizing emissivity now matches the growth of the galaxy UV luminosity density. The Astrophysical Journal 811 (2), 140.

Broadhurst, T.J., Taylor, A.N., Peacock, J.A., 1995. Mapping cluster mass distributions via gravitational lensing of background galaxies. The Astrophysical Journal 438, 49–61.

Buchmueller, O., et al., 2012. Higgs and supersymmetry. European Physical Journal C 72. 2020.

Burles, S., Tytler, D., 1998. The Deuterium Abundance toward Q1937-1009. The Astrophysical Journal 499, 699–712.

Clifton, T., Ferreira, P.G., Padilla, A., Skordis, C., 2012. Modified gravity and cosmology. Physics Reports 513 (1–3), 1–189.

Cooke, R.J., Pettini, M., Steidel, C.C., 2018. One percent determination of the primordial deuterium abundance. The Astrophysical Journal 855 (2), 102.

Cooray, A., Sheth, R.K., 2002. Halo models of large scale structure. Physics Reports 372, 1–129.

Cowsik, R., McClelland, J., 1972. An upper limit on the neutrino rest mass. Physical Review Letters 29, 669–670.

Desjacques, V., Jeong, D., Schmidt, F., 2018. Large-scale galaxy bias. Physics Reports 733, 1–193.

Dodelson, S., 2003. Coherent phase argument for inflation. AIP Conference Proceedings 689 (1), 184–196.

Dyson, F.W., Eddington, A.S., Davidson, C., 1920. A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of May 29, 1919. Philosophical Transactions of the Royal Society of London Series A 220, 291–333.

Elvin-Poole, J., et al., DES Collaboration, 2018. Dark energy survey year 1 results: galaxy clustering for combined probes. Physical Review D 98 (4), 042006.

Fixsen, D.J., 2009. The temperature of the cosmic microwave background. The Astrophysical Journal 707 (2), 916–920.

Fixsen, D.J., et al., 1996. The cosmic microwave background spectrum from the full COBE FIRAS data set. The Astrophysical Journal 473, 576.

Freedman, W.L., et al., 2001. Final results from the Hubble space telescope key project to measure the Hubble constant. The Astrophysical Journal 553, 47–72.

Frieman, J.A., Turner, M.S., Huterer, D., 2008. Dark energy and the accelerating universe. Annual Review of Astronomy and Astrophysics 46, 385–432.

Fukuda, Y., et al., 1998. Evidence for oscillation of atmospheric neutrinos. Physical Review Letters 81, 1562–1567.

Fukugita, M., Hogan, C.J., Peebles, P.J.E., 1998. The cosmic baryon budget. The Astrophysical Journal 503, 518–530.

Gershtein, S.S., Zel'dovich, Y.B., 1966. Rest mass of muonic neutrino and cosmology. JETP Letters 4, 120–122. 58 (1966).

Gil-Marín, H., et al., 2016. The clustering of galaxies in the SDSS-III baryon oscillation spectroscopic survey: RSD measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies. Monthly Notices of the Royal Astronomical Society 460, 4188–4209.

Gunn, J.E., et al., 1978. Some astrophysical consequences of the existence of a heavy stable neutral lepton. The Astrophysical Journal 223, 1015–1031.

Guth, A.H., 1981. Inflationary universe: a possible solution to the horizon and flatness problems. Physical Review D 23, 347–356.

Heitmann, K., Lawrence, E., Kwan, J., 2014. The coyote universe extended: precision emulation of the matter power spectrum. The Astrophysical Journal 780, 111.

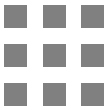
Hezaveh, Y.D., et al., 2016. Detection of lensing substructure using ALMA observations of the dusty galaxy SDP.81. The Astrophysical Journal 823 (1), 37.

Hu, W., Okamoto, T., 2002. Mass reconstruction with cosmic microwave background polarization. The Astrophysical Journal 574, 566–574.

Hu, W., Sugiyama, N., 1995. Anisotropies in the cosmic microwave background: an analytic approach. The Astrophysical Journal 444, 489–506.
Hu, W., Sugiyama, N., 1996. Small-scale cosmological perturbations: an analytic approach. The Astrophysical Journal 471, 542.
Hu, W., 2001. Mapping the dark matter through the CMB damping tail. The Astrophysical Journal 557, L79–L83.
Hu, W., White, M., 1997. A CMB polarization primer. New Astronomy 2, 323–344.
Hubble, E., 1929. A relation between distance and radial velocity among extra-galactic nebulae. Proceedings of the National Academy of Science 15 (3), 168–173.
Joyce, A., Lombriser, L., Schmidt, F., 2016. Dark energy versus modified gravity. Annual Review of Nuclear and Particle Science 66, 95–122.
Kaiser, N., 1984. On the spatial correlations of Abell clusters. The Astrophysical Journal Letters 284, L9–L12.
Kaiser, N., 1987. Clustering in real space and in redshift space. Monthly Notices of the Royal Astronomical Society (ISSN 0035-8711) 227, 1–21.
Kochanek, C.S., 1996. Is there a cosmological constant? The Astrophysical Journal 466, 638.
Kodama, H., Sasaki, M., 1984. Cosmological perturbation theory. Progress of Theoretical Physics. Supplement 78, 1.
Landau, L.D., Lifshitz, E.M., 1965. Quantum mechanics.
Le Tiec, A., Novak, J., 2017. An overview of gravitational waves: theory, sources and detection. In: Plagnol, E., Auger, G. (Eds.), Theory of Gravitational Waves, pp. 1–41.
Lewis, A., Challinor, A., Lasenby, A., 2000. Efficient computation of cosmic microwave background anisotropies in closed Friedmann-Robertson-Walker models. The Astrophysical Journal 538, 473–476.
Liddle, A.R., Lyth, D.H., 2000. Cosmological Inflation and Large-Scale Structure.
Linde, A.D., 1982. A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. Physics Letters B 108, 389–393.
Louis, T., et al., 2017. The Atacama Cosmology Telescope: two-season ACTPol spectra and parameters. Journal of Cosmology and Astroparticle Physics 1706 (06), 031.
Ma, C.-P., Bertschinger, E., 1995. Cosmological perturbation theory in the synchronous and conformal Newtonian gauges. The Astrophysical Journal 455, 7.
Makino, N., Sasaki, M., Suto, Y., 1992. Analytic approach to the perturbative expansion of nonlinear gravitational fluctuations in cosmological density and velocity fields. Physical Review D 46, 585–602.
Mantz, A.B., et al., 2014. Cosmology and astrophysics from relaxed galaxy clusters – II. Cosmological constraints. Monthly Notices of the Royal Astronomical Society 440 (3), 2077–2098.
Martin, J., 2012. Everything you always wanted to know about the cosmological constant problem (but were afraid to ask). Comptes Rendus. Physique 13, 566–665.
Martin, J., Ringeval, C., Vennin, V., 2014. Encyclopædia inflationaris. Physics of the Dark Universe 5–6, 75–235.
Meszaros, P., 1974. The behaviour of point masses in an expanding cosmological substratum. Astronomy & Astrophysics 37, 225–228.
Mo, H., van den Bosch, F.C., White, S., 2010. Galaxy Formation and Evolution.
Moessner, R., Jain, B., 1998. Angular cross-correlation of galaxies – a probe of gravitational lensing by large-scale structure. Monthly Notices of the Royal Astronomical Society 294, L18–L24.
Mortonson, M.J., Weinberg, D.H., White, M., 2014. Dark energy: a short review. Review of Particle Physics.
Mukhanov, V., 2005. Physical Foundations of Cosmology.
Mukhanov, V.F., Feldman, H.A., Brandenberger, R.H., 1992. Theory of cosmological perturbations. Part 1. Classical perturbations. Part 2. Quantum theory of perturbations. Part 3. Extensions. Physics Reports 215, 203–333.
Navarro, J.F., Frenk, C.S., White, S.D.M., 1997. A universal density profile from hierarchical clustering. The Astrophysical Journal 490, 493–508.
Olive, K.A., 2000. Big Bang nucleosynthesis. Nuclear Physics. B, Proceedings Supplement 80, 79–93.
Partridge, R.B., 2007. 3K: The Cosmic Microwave Background Radiation.
Peebles, P.J.E., 1968. Recombination of the primeval plasma. The Astrophysical Journal 153, 1.
Perlmutter, S., et al., 1999. Measurements of Omega and Lambda from 42 high redshift supernovae. The Astrophysical Journal 517, 565–586.

Planck Collaboration, 2018a. Planck 2018 results. I. Overview and the cosmological legacy of Planck. arXiv: 1807.06205.
Planck Collaboration, 2018b. Planck 2018 results. VI. Cosmological parameters. arXiv:1807.06209.
Polnarev, A.G., 1985. Polarization and anisotropy induced in the microwave background by cosmological gravitational waves. Soviet Astronomy 29, 607–613.
Pospelov, M., Pradler, J., 2010. Big Bang nucleosynthesis as a probe of new physics. Annual Review of Nuclear and Particle Science 60, 539–568.
Prat, J., et al., DES Collaboration, 2018. Dark energy survey year 1 results: galaxy-galaxy lensing. Physical Review D 98, 042005.
Press, W.H., Schechter, P., 1974. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. The Astrophysical Journal 187, 425–438.
Renn, J., Sauer, T., Stachel, J., 1997. The origin of gravitational lensing: a postscript to Einstein's 1936 science paper. Science 275, 184–186.
Riess, A.G., et al., 1998. Observational evidence from supernovae for an accelerating universe and a cosmological constant. The Astronomical Journal 116, 1009–1038.
Sachs, R.K., Wolfe, A.M., 1967. Perturbations of a cosmological model and angular variations of the microwave background. The Astrophysical Journal 147, 73.
Sato, K., 1981. First-order phase transition of a vacuum and the expansion of the universe. Monthly Notices of the Royal Astronomical Society 195, 467–479.
Schmidt, F., 2016. Towards a self-consistent halo model for the nonlinear large-scale structure. Physical Review D 93 (6), 063512.
Schumann, M., 2012. Dark Matter Search with liquid Noble Gases.
Scolnic, D.M., et al., 2018. The complete light-curve sample of spectroscopically confirmed SNe Ia from Pan-STARRS1 and cosmological constraints from the combined Pantheon sample. The Astrophysical Journal 859, 101.
Seljak, U., 1994. A two-fluid approximation for calculating the cosmic microwave background anisotropies. The Astrophysical Journal Letters 435, L87–L90.
Shull, J.M., Smith, B.D., Danforth, C.W., 2012. The baryon census in a multiphase intergalactic medium: 30% of the baryons may still be missing. The Astrophysical Journal 759, 23.
Smoot, G.F., et al., 1992. Structure in the COBE differential microwave radiometer first-year maps. The Astrophysical Journal Letters 396, L1–L5.
Springel, V., et al., 2005. Simulations of the formation, evolution and clustering of galaxies and quasars. Nature (London) 435, 629–636.
Srednicki, M., 2007. Quantum Field Theory. Cambridge Univ. Press, Cambridge.
Starobinsky, A.A., 1982. Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations. Physics Letters B 117, 175–178.
Steigman, G., 2007. Primordial nucleosynthesis in the precision cosmology era. Annual Review of Nuclear and Particle Science 57, 463–491.
Szalay, A.S., Marx, G., 1976. Neutrino rest mass from cosmology. Astronomy & Astrophysics 49 (3), 437–441.
Tanabashi, M., et al., 2018. Review of particle physics. Physical Review D 98, 030001.
Tinker, J., et al., 2008. Toward a halo mass function for precision cosmology: the limits of universality. The Astrophysical Journal 688, 709–728.
Tisserand, P., et al., 2007. Limits on the macho content of the galactic halo from the EROS-2 survey of the Magellanic Clouds. Astronomy & Astrophysics 469, 387–404.
Troxel, M.A., et al., DES Collaboration, 2018. Dark energy survey year 1 results: cosmological constraints from cosmic shear. Physical Review D 98, 043528.
Tulin, S., Yu, H.-B., 2018. Dark matter self-interactions and small scale structure. Physics Reports 730, 1–57.
Tyson, J.A., Valdes, F., Wenk, R.A., 1990. Detection of systematic gravitational lens galaxy image alignments – mapping dark matter in galaxy clusters. The Astrophysical Journal Letters 349, L1–L4.
Vegetti, S., et al., 2012. Gravitational detection of a low-mass dark satellite galaxy at cosmological distance. Nature (London) 481 (7381), 341–343.
White, S.D.M., Frenk, C.S., Davis, M., 1983. Clustering in a neutrino-dominated universe. The Astrophysical Journal Letters 274, L1–L5.
Wong, Kenneth C., et al., 2019. H0LiCOW XIII. A 2.4% measurement of $H_0$ from lensed quasars: $5.3\sigma$ tension between early and late-Universe probes.

Zaldarriaga, M., Harari, D.D., 1995. Analytic approach to the polarization of the cosmic microwave background in flat and open universes. Physical Review D 52, 3276–3287.
Zel'dovich, Y.B., Sunyaev, R.A., 1969. The interaction of matter and radiation in a hot-model universe. Astrophysics and Space Science 4, 301–316.
Zwicky, F., 1933. Die Rotverschiebung von extragalaktischen Nebeln. Helvetica Physica Acta 6, 110–127.



# Index


**A**

Abundance, 9, 85, 100, 360
  baryons, 94
  dark matter, 103, 104, 106
  deuterium, 8, 94, 95
  equilibrium, 100
  halo, 368
  light element, 1, 8, 94
Abundance cluster, 366
Abundance matching technique, 363
Accelerated expansion, 47, 50, 159, 162, 163, 352
  transitory epoch, 163
Acoustic oscillations, 238
Acoustic peaks, 257
Adiabatic perturbations, 174, 177, 183, 185, 191, 266
Alcock–Paczyński, 305
Alignments
  intrinsic, 391
Amplitude, 232, 262
Anisotropies
  CMB, 10, 17, 43, 47, 78, 96, 98, 226, 232, 265, 415
    temperature, 265
  large-scale, 237
  polarization, 1, 271
  primordial, 262, 318
  spectrum, 231, 232, 244, 254, 257, 285
  temperature, 1, 251, 256, 268, 271, 278, 285, 287–289, 407
  tensor, 269
Annihilation operator, 169
Argument
  peak-background split, 360
Atomic number, 89

**B**

Background
  cosmology, 397
  FLRW metric, 137
  galaxies, 358, 374–376, 396, 398, 399
  metric, 136, 168
  universe, 14, 34, 39, 58, 61, 138, 196, 297, 352
  unperturbed, 175
Baryon acoustic oscillation (BAO), 49, 223, 295, 305
  feature, 320
Baryons, 41, 222, 263
  abundance, 94
  Boltzmann equation for, 126
  density, 8, 9, 42, 95, 96, 236, 241, 260, 264
  energy density, 129
  evolution, 224
  overdensity, 185
  velocity, 132, 317
Bayes' theorem, 403
Bias, 359
Big Bang Nucleosynthesis (BBN), 8, 42, 88
Bispectrum, 341
Boltzmann distribution, 86
Boltzmann equation, 62, 79, 85
  collisionless, 64, 329, 345
  dark matter, 127
  for baryons, 126
  for cold dark matter, 122
  for dark matter, 111
  for neutrinos, 129, 130
  for particles, 63
  for photons, 112, 119, 122, 128, 130, 132, 268, 316, 321
  for radiation, 77, 78
  in an expanding universe, 65
  tightly-coupled limit, 238

Boltzmann hierarchy, 200
Boltzmann solutions, 283
Bulk velocity, 117–119

C
Calibration, 409
CAMB, 195, 226
Causal horizon, 183
CDM, 15
  densities, 263
  overdensity, 197
  perturbations, 201
Chain, 425
Christoffel symbol, 27, 28, 35, 59, 81, 82, 141, 148, 149
  for tensor perturbations, 148
CLASS, 195, 226
Cluster mass, 358
  calibration, 396
Clustering
  projected, 320
CMB
  anisotropies, 10, 17, 43, 47, 78, 96, 98, 226, 231, 232, 259, 265, 415
  lensing, 315, 381–384
    deflections, 398
  photons, 237, 262, 287, 291, 296, 315, 318–320
  polarization, 117, 260, 271, 278, 291
  power spectrum, 13, 231, 247, 255, 413
  radiation, 251
  temperature, 12, 13, 241, 260, 295, 315, 316, 384, 406, 409, 411, 412
    anisotropies, 265
    lensing, 384
    perturbations, 316
    power spectrum, 260
Cold dark matter (CDM), 15, 40, 103, 106, 122, 123, 125, 222, 224, 265, 376
  Boltzmann equation for, 122
  halos, 374
Collision term, 63, 69, 70, 80, 115, 116, 128
  vanishes, 113, 317
Collisionless
  Boltzmann equation, 64, 329, 345
    for massive particles, 78, 122, 129
    for photons, 112
    for radiation, 77
  dark matter, 375
  neutrinos, 146
  particles, 82, 342
Collisions, 68
Comoving
  distance, 1, 7, 30, 31, 158, 235, 236, 398
  gauge, 183
  grid, 30, 33, 159, 295
  horizon, 158, 159, 163, 184, 190, 202
  Hubble radius, 159, 161, 162, 227, 228
Compression, 428
Compton drag, 128
Compton scattering, 95, 99, 114, 238, 271, 276, 278
  polarization dependence, 278
Concentration, 350
Concordance cosmology, 7, 186, 305, 347
  Euclidean, 49, 50
  fiducial, 31, 54
Concordance model, 1, 9, 42, 43, 224, 260
  cosmology, 15
Constant normalization factor, 403
Constraint equations, 147
Convergence, 396
Conversion factor, 180
Correlation function, 315, 321, 361, 371
  galaxy, 430
  shear, 391
Correlations
  angular, 311
Cosmic
  inventory, 40
  neutrinos, 43, 105
  plasma, 6, 10, 44, 88, 100, 106, 109
  variance, 253, 262, 265, 288, 291, 407, 409, 416, 417, 422
    error, 263
Cosmic microwave background (CMB), 4, 9, 40
Cosmological
  constant, 4, 15, 18, 49, 55, 58, 260
  perturbations, 137, 138
    evolution, 196

redshift, 295, 297, 307, 378
Cosmological parameters, 259
Cosmology
background, 397
concordance model, 15
Euclidean, 71, 218, 221
fiducial, 156, 187, 194, 202, 207, 233, 249, 268, 296, 298
Covariances, 428
Covariant derivative, 34–36
Creation operator, 169
Cross-component, 392
Cross-correlations
shear, 393
Curvature perturbation, 177, 182, 183, 187, 232, 237, 262
power spectrum, 254, 256
primordial, 254
variance, 187, 255

**D**

Dark energy, 4, 15, 47, 50, 61, 225
Dark Energy Survey (DES), 315
Dark Energy Survey Instrument (DESI), 401
Dark matter, 9, 42, 99
abundance, 103, 104, 106
Boltzmann equation, 111, 127
collisionless, 375
density, 104, 202
evolution, 122, 213, 214
halos, 326, 346, 347, 349, 350, 364, 367, 375
linear evolution, 326
mass, 109
overdensity, 17, 185, 200, 238
particles, 15, 104, 105, 122, 184, 344, 364
perturbations, 214, 215, 217, 218
Decomposition
scalar–vector–tensor, 135
theorem, 137
Decoupled relativistic particles, 88
Decoupling epoch, 95, 98
Degeneracy factor, 63, 80, 87, 144
neutrinos, 44
Density field, 200, 302, 321, 333, 371
nonlinear, 397
Density parameter, 40
Density perturbations, 174
primordial, 174
Detection
direct, 103, 105
indirect, 103, 104
Deuterium, 42, 91, 95
abundance, 8, 94, 95
mass, 89
nucleus, 89
Diffusion damping, 244
Dipole, 128, 130, 131, 195, 200, 201, 227, 239, 257, 258, 286
moments, 119, 185
pattern, 277, 278
Distance modulus, 48
Distant galaxies, 1, 4, 6, 31, 312, 373, 374, 376, 388
Distortion tensor, 381
Distortions
Alcock–Paczyński, 320
redshift-space, 301, 320
Distribution function, 37, 64, 80, 83, 329
neutrinos, 83
perturbations, 112
photons, 114, 155, 156, 377, 397

**E**

Effect
Alcock–Paczyński (AP), 309
fractional, 297
Sunyaev–Zel'dovich, 315, 321
Effective fluid, 342
Einstein equations, 57, 79, 143
for gravity, 73
for scalar perturbations, 141
for tensor perturbations, 150
tensor, 151
Einstein–Boltzmann equations
at early times, 183
Electrons
free, 9, 95, 111, 237, 246, 262, 287
mass, 44, 93, 115, 127
temperature, 316
velocity, 119, 129

Energy
density
in gravitational waves redshifts, 152
in neutrinos, 224
evolution, 34
integral, 70
Energy-momentum tensor, 34, 80
Epoch
of inflation, 16
of matter–radiation equality, 47
recombination, 231, 262, 291
Equation of state, 37
negative, 47
Equilibrium
abundance, 100, 101
chemical, 87, 115
distribution, 65, 77, 129, 318, 426
kinetic, 115
nuclear statistical, 87
Estimator, 404
Euclidean
concordance cosmology, 49, 50
concordance universe, 54
cosmology, 71, 218, 221
FLRW metric, 28, 81
FLRW universe, 81
models, 268
space, 27
universe, 2, 4, 15, 26, 49, 51, 59, 79, 81, 136, 261–263
Evolution
baryons, 224
dark matter, 122, 213, 214
equations, 85, 101, 143–145, 151, 200, 203
for perturbations, 111
linear, 340, 360
nonlinear, 329, 340, 341, 366
perturbations, 31, 52, 145, 197, 265
stages, 196
sub-horizon, 202, 214
Expanding space, 21
Expansion
history, 57, 224, 295, 296, 346, 391, 393
rate, 4, 76, 108, 370
Expectation value, 422
Exponential factor, 94, 255

**F**

Fast Fourier Transform (FFT), 346
Fiducial
concordance cosmology, 31, 54
cosmology, 156, 187, 194, 202, 207, 233, 249, 268, 296, 298
Euclidean concordance cosmology, 54
Euclidean ΛCDM cosmology, 218
Euclidean ΛCDM model, 256
Euclidean ΛCDM prediction, 265
ΛCDM cosmology, 390, 398
Fisher information, 422
Fisher matrix, 421, 422
Flatness problem, 189
FLRW metric, 28–30, 34, 36, 59, 71, 72, 79, 137
background, 137
Euclidean, 28, 81
FLRW universe, 157, 352
Euclidean, 81
homogeneous, 147
perturbed, 75
Fluid
effective, 342
Forecasting, 421
Fourier modes, 122, 135, 231, 255, 257, 304, 417, 418, 420
Fractional
overdensity, 13, 124
perturbations, 183
temperature perturbation, 112
Free
electrons, 9, 95, 111, 237, 246, 262, 287
particles, 26, 64
protons, 98, 106, 133
Free streaming, 119, 247
scale, 224
Friedmann equation
first, 61
Fundamental particles, 14, 106

**G**

Galaxy
bias, 304, 420, 428

bispectrum, 363
clustering, 43, 195, 295, 296, 320, 359, 413
clusters, 321, 350, 355, 356, 374
correlation function, 430
density, 300, 372, 391, 394, 406, 412, 417, 420
field, 295, 299, 311
perturbation, 361
distribution, 12, 388, 407, 412
ellipticities, 385, 391, 407
ellipticity power spectrum, 390
image, 374, 377, 384–387
number counts, 43, 376
overdensity, 300, 359, 411, 417
peculiar velocities, 295, 320
positions, 299, 302, 305, 307, 311, 384
power spectrum, 12, 295, 302, 304, 308–310, 320, 323, 409, 417–421, 430
redshift surveys, 295, 296, 309, 310, 401, 411, 418
redshifts, 295, 389, 395
shape correlations, 396
shapes, 43, 373, 381, 384, 391
source, 391, 394, 395, 397, 429
statistics, 299
surveys, 300, 304, 320, 358, 407
velocities, 43, 295–298, 359, 363
Gas
pressure, 319
temperature, 318
ranges, 315
velocity, 320
Gauge, 137
synchronous-comoving, 329
Geodesic equation, 26–28, 30, 65, 73, 74, 77, 79, 345, 363, 378, 379
Gradient
of a 3-scalar function, 136
Gravitational
lensing, 10, 43, 77, 373, 381, 396
physics, 61, 136
potential, 196, 201, 209, 214, 215, 223, 317, 347, 349, 391
energy, 357
perturbations, 211
wave production, 167
waves, 58, 147, 151, 154, 167, 170, 172, 288, 292
power spectrum, 170, 173
Gravity, 57, 71, 79, 111, 184, 305, 325
Growth factor, 197, 219–221, 225, 226
Growth rate, 221

**H**

Halo, 325
abundance, 368
density perturbation, 360
mass, 349, 350, 372
Milky Way, 351
mass function, 350, 354, 365, 368
model, 349, 364
overdensity, 360, 372
profiles, 349
Halo Occupation Distribution (HOD), 363
Harmonic oscillator, 168
Harmonic potential, 63, 64, 82
Headless vector, 271
Helium, 91, 94, 96, 97, 126, 246
atoms, 96, 129
mass fraction, 95
neutral, 96
nucleus, 246
Homogeneous universe, 70, 325, 350
Horizon
comoving, 158, 159, 184, 190, 202
crossing, 177, 179, 188, 191, 203, 207, 209
epochs, 196, 197
entry, 202, 212
problem, 157, 159, 162, 163, 167
solution, 157
Hubble diagram, 6, 7
Hubble expansion, 75, 295, 345, 352
Hubble rate, 3, 61, 102, 108, 160, 162, 164, 186, 187, 225, 308
Hubble-Lemaître law, 7
Hubble's constant, 3

**I**

Incoming
photon polarization, 278
photons, 251

radiation, 278, 279, 283
Inflation, 16, 159, 163
Inflationary
gravitational waves, 290
models, 167, 173, 181, 188, 189, 193
paradigm, 262
perturbations, 182
predictions, 167
scenario, 168, 262, 291
Inhomogeneities, 10, 12, 71, 112, 195, 200, 236, 247, 260, 373
Integral
energy, 70
term, 244
Integrated Sachs–Wolfe effect, 236
Intrinsic alignments, 391
Inverse-variance weighting, 404
Isotopes, 89
Isotropic radiation, 277
Isotropic smooth universe, 34

**J**
Jacobian factor, 386

**K**
Kinetic energy, 38
density, 164

**L**
Lagrangian
approach, 370
radius, 351, 368
Large scales, 203
Large-scale
anisotropies, 237
Large-Scale Structure (LSS), 11
Lens galaxies, 394–397
redshift range, 395
Lensing, 195, 373, 374, 393, 396, 397, 429
CMB, 315, 381–384
temperature, 384
deflection angle, 384
deflections, 382
power spectrum, 382
distortions, 373
effect, 383, 384, 394, 399
field, 383
gravitational, 10, 43, 77, 373, 381, 396
kernel, 390, 395–397
potential, 380–384, 389, 397–399, 411
shear, 393
signal, 358, 388
statistics, 388, 391
weak, 396
Likelihood function, 402, 424
Limber approximation, 314
Limit
nonrelativistic, 67
relativistic, 67
Linear
bias relation, 302
evolution, 340, 360
Logarithmic derivative, 188, 221
Loop, 340

**M**
Macroscopic
pressure, 38
quantity, 37
Magnification, 374, 387
bias, 399
Malmquist bias, 358
Map, 23
Mapmaking, 409
Markov Chain Monte Carlo (MCMC), 402, 425
Mass
bin, 350
dark matter, 109
density, 326, 330, 354
deuterium, 89
difference, 91, 96
distribution, 373, 395
electrons, 44, 93, 115, 127
fraction, 95
function, 355, 360, 364, 366
halo, 350, 354, 365, 368
halo, 349, 350, 372
Milky Way, 351
map, 376, 384
number, 89, 91

particles, 36
protons, 127, 357
range, 100
solar, 5, 106
Mass–observable relation, 358
Mass-shell constraint, 65
Massive
bosons, 14
bound structures, 367
clusters, 358
abundance, 356
collapsed objects, 358
compact halo objects, 374
dark matter
halos, 326
particles, 105
elliptical galaxies, 363
fermions, 14
galaxy clusters, 43, 296, 316, 319, 375
halos, 347, 350, 351, 353–355
lenses, 374
neutrinos, 45, 125, 130, 224–226, 228, 368
energy density, 46
particles, 29, 53, 54, 74, 78, 81, 83, 160, 363
stars, 347
Massless
bosons, 44
fermions, 44
field, 176
gravitons, 176
neutrinos, 52, 55, 83, 145
particles, 25, 29, 30, 44, 52–54, 77, 160
photon, 40
Metric, 22
background, 136, 168
convention, 23
perturbations, 71, 74, 111, 135, 137, 140, 147, 177, 273
variance, 328
perturbed, 73, 79, 80, 83, 135–138, 140, 141
signature, 164
tensor, 25, 60
Microlensing, 374
Minkowski metric, 26
Minkowski spacetime, 25
Momentum
comoving, 63
physical, 63
Monopole, 118, 131, 195, 200, 201, 227, 239, 255, 258, 265, 267, 286, 292
equation, 227
perturbation, 130
photons, 232
spectrum, 257
Multiplication factor, 250
Multipole moments, 118

**N**

Nearby electrons, 119
Nearby universe, 320
Negative pressure, 47, 53, 163, 165
Neutral
helium, 96
particles, 106
stable particle, 106
Neutrinos, 43
Boltzmann equation for, 129, 130
collisionless, 146
cosmic, 43, 105
degeneracy factor, 44
density, 45
distribution function, 83
energy density, 44, 45, 224
generation, 53
masses, 14, 45, 52, 83, 130, 145, 224, 260
massive, 45, 125, 130, 224–226, 228, 368
massless, 52, 83, 145
perturbations, 131
quadrupole, 206
sterile, 44
temperature, 44, 45, 52, 183
thermal, 83
Neutron abundance, 91, 93–95
Newtonian gravity, 58, 328, 332
Newtonian physics, 76, 77
Next-to-Leading Order (NLO), 338
NFW profile, 350
Noise covariance matrix, 411, 429

Nonlinear
density field, 397
evolution, 329, 340, 341, 366
gravitational evolution, 342
order, 329, 332, 334
overdensity, 371
perturbations, 200
Nonnegligible
peculiar velocities, 295
quadrupole, 244
Nonrelativistic
matter, 3, 4, 16, 36, 40, 53, 78, 163
energy density, 4
temperature, 83
particles, 38, 62, 63, 76, 77, 88, 109, 146
Number density, 354
Nutshell history, 1

**O**

Operator
annihilation, 169
creation, 169
Optical depth, 262
Optimal quadratic estimator, 384
Oscillations
acoustic, 238
Outgoing
photons, 117, 278
polarization, 278, 283, 289
radiation, 278, 292
Overdensity, 130, 168, 196, 234, 238, 303, 304, 352, 357, 380, 395
baryons, 185
dark matter, 17, 185, 200, 238
equations, 183
fractional, 124
galaxy, 300, 359, 411, 417
halo, 360, 372
in redshift space, 303
nonlinear, 371
radiation, 185

**P**

Parameter degeneracy, 52
Parameters
cosmological, 259
Partial derivatives, 35, 64, 180
Particles
Boltzmann equation, 63
collisionless, 82, 342
dark matter, 15, 104, 105, 122, 184, 344, 364
free, 26, 64
interaction, 68–70
mass, 36
massive, 29, 53, 54, 74, 78, 81, 83, 160, 363
massless, 25, 29, 30, 44, 52–54, 77, 160
neutral, 106
nonrelativistic, 38, 62, 63, 76, 77, 88, 109, 146
physics, 1, 14–16, 46, 106, 163, 164
positions, 345, 346
trajectory, 370
velocity, 123, 345, 346
Peak-background split argument, 360
Peaks
acoustic, 257
Peculiar velocity, 76, 297, 301, 303, 308, 309
Perturbations
CMB temperature, 316
cosmological, 137, 138
dark matter, 215, 217, 218
distribution function, 112
evolution, 31, 52, 145, 197, 265
fractional, 183
in Fourier space, 133
inflationary, 182
isocurvature, 174, 266
metric, 71, 74, 111, 135, 137, 140, 147, 167, 177, 273
neutrinos, 131
photons, 118, 131, 231, 237
pressure, 196, 319
primordial, 14, 17, 186, 231, 237, 262, 265
radiation, 195, 214, 215, 222, 244
relative, 266
scalar, 167, 173
scalar field, 175
super-horizon, 177
tensor, 71, 147, 149, 170, 176, 179, 288, 289

theory, 137, 330, 338
variables, 124, 144, 183
variance, 171
velocity, 174
Perturbed
Boltzmann equations, 130, 135
metric, 73, 79, 80, 83, 135–138, 140, 141
spacetime, 71
universe, 71, 73, 77, 79, 80
Photometric
galaxy, 320
redshifts, 311, 393, 395
Photons, 40
Boltzmann equation, 112, 119, 122, 128, 130, 132, 268, 316, 321
CMB, 237, 262, 287, 291, 296, 315, 318–320
collisionless Boltzmann equation for, 112
density, 41, 46
diffusion, 240, 252
direction, 118, 121, 248
distribution, 115, 200, 243, 254, 278, 286, 288, 318
distribution function, 112, 114, 155, 156, 377, 397
energies, 115, 116
energy density, 44, 107, 144
geodesics, 373, 376
incoming, 251
moments, 113, 115, 128, 193, 200, 227, 241, 247, 288, 379
monopole, 232
number, 39, 40
number density, 55
outgoing, 117, 278
path, 378–380
perturbation variable, 131
perturbations, 118, 131, 231, 237
polarization states, 278
propagation, 121
quadrupole, 146, 193
temperature, 52, 109, 114, 242, 317, 318
perturbation, 265
Physics
gravitational, 61, 136
particles, 1, 14–16, 46, 106, 163, 164
Planck
epoch, 189
mass, 173, 194
Pointing matrix, 409
Polarization, 271
anisotropies, 1, 271
axes, 278
CMB, 117, 260, 271, 278, 291
dependence, 117
field, 131, 132, 288
from Compton scattering, 275
outgoing, 278, 283, 289
pattern, 271, 274, 275, 283, 289, 290, 292, 385
power spectra, 285
power spectrum, 285, 286
radiation, 272
signal, 288
single plane wave, 278
spectrum, 285, 286
strength, 283, 288, 292
tensor, 272, 274, 373, 385
Polarized radiation, 276, 278
Compton scattering, 271
Posterior, 403
marginalized, 407
Postulated epoch, 159
Potential
energy density, 164
harmonic, 63, 64, 82
Potential perturbations, 225
gravitational, 211
Power spectra
polarization, 285
Power spectrum, 12, 13, 171, 199, 337, 391, 419
angular, 251, 285
CMB, 13, 231, 247, 255, 413
temperature, 260
galaxy, 12, 295, 302, 304, 308–310, 320, 323, 409, 417–421, 430
polarization, 285, 286
primordial, 198
shear, 388
Prefactor, 185, 218, 239, 249, 360, 417
Pressure
forces, 325, 344, 347, 363

gas, 319
gradient, 196
map, 321
perturbations, 196, 319
radiation, 178, 214
Primordial
abundances, 8
anisotropies, 262, 318
CMB anisotropies, 321
curvature perturbation, 254
density perturbations, 174
gravitational waves, 173, 268
helium abundance, 95
perturbations, 14, 17, 186, 231, 237, 262, 265
power spectrum, 198
temperature power spectra, 382
tensor modes, 187, 289
tensor perturbations, 171
Protons, 9, 91
free, 106, 133
mass, 127, 357

**Q**

Quadrupole
moment, 146, 201, 278, 280, 285
nonnegligible, 244
pattern, 290
photons, 146, 193
temperature, 269, 287
temperature field, 291
Quintessence, 163

**R**

Radiation
Boltzmann equation for, 77, 78
CMB, 251
collisionless Boltzmann equation for, 77
components, 151, 196
density, 53
domination, 3, 146, 162, 184, 186, 187, 192, 196, 202, 213, 225
energy density, 37, 41, 53, 192
epochs, 199, 206
equations, 207, 210
equilibrium distribution, 78
era, 101, 102, 166, 206, 209, 212–214, 217, 218
field, 222, 271, 278
incident, 271, 277, 278
incoming, 278, 279, 283
outgoing, 278, 292
overdensity, 185
perturbations, 195, 214, 215, 222, 244
polarization, 272
pressure, 178, 214
Radius
comoving, 351
Lagrangian, 351
Recombination, 95
epoch, 231, 262, 291
neglecting, 246
rate, 97
redshift, 97
Redshift
cosmological, 295, 297, 307, 378
distribution, 391, 397
interval, 49
recombination, 97
relation, 4
slice, 300, 307, 315
space, 299, 301, 321
surveys, 295, 296, 310
Redshift-space distortions (RSDs), 295, 301
Reionization, 99, 260, 262, 263, 268, 287, 288, 319
bump, 288
Relative
perturbations, 266
velocity, 223, 224
Ricci scalar, 58, 60, 141, 150
Ricci tensor, 58, 59, 141, 149, 150

**S**

Saha equation, 87, 96
Sample variance, 407
Scalar, 35
perturbations, 173
Ricci, 58, 60, 141, 150
Scalar field
perturbations, 175

Scales
factor, 1, 32, 40, 47, 71, 76, 160, 179, 216, 228, 266
large, 203
radius, 350
small, 209
Scattering
Thomson, 113
Second Friedmann equation, 81
Shear, 396
correlation function, 391
cross-correlations, 393
tangential, 392
Shell crossing, 343
Signal
Sunyaev–Zel'dovich (SZ), 296
Simulations, 343
Single
distribution function, 272
fluid, 235
massive species, 228
mode, 274
parameter, 422, 426, 431
particle, 29, 87, 165, 169
photon, 235
pixel, 409, 411
proton, 89
Sloan Digital Sky Survey (SDSS), 296
Slow-roll models, 166
Small scales, 209
Solar masses, 5, 106
Solutions
Boltzmann, 283
super-horizon, 203
tightly-coupled, 242
Sound horizon, 223, 242, 264, 266, 268
Sound speed, 191
Spacetime
metric, 22
perturbations, 73, 78, 325, 328, 329, 373
perturbed, 71
Spatially flat slicing, 174, 181
Spectral index, 262
Spectrum
scale-free, 186
scale-invariant, 186
Spherical collapse threshold, 353
Spherical overdensity, 348
Statistics
weak-lensing, 388
Stress tensor, 331
Structure, 10
universe, 10
Sub-horizon
evolution, 214
Sunyaev–Zel'dovich effect, 315, 321
Super-horizon
perturbations, 177
regime, 202
solutions, 203
Symmetric
3-tensor, 136
Christoffel symbol, 72
tensor, 137, 148
System
Vlasov–Poisson, 329

**T**

Technique
abundance matching, 363
Temperature
anisotropies, 1, 251, 256, 268, 271, 278, 285, 287–289, 407
anisotropy spectrum, 285
CMB, 12, 13, 241, 260, 295, 315, 316, 384, 406, 409, 411, 412
derivative, 114
distribution, 132, 252, 382
electrons, 316
field, 131, 132, 236, 251, 252
fluctuations, 10, 157, 187, 413
gas, 318
neutrinos, 44, 45, 52, 183
perturbation, 112, 117, 118, 129, 252, 413, 429
photons, 52, 109, 114, 242, 317
primordial, 382
quadrupole, 269, 287, 293
quadrupole pattern, 290
spectrum, 286

Tensor
  anisotropies, 269
  degrees of freedom, 137
  Einstein equations, 151
  fluctuations, 167
  metric, 25, 60
  metric perturbations, 273, 288
  modes, 150, 151, 154, 168, 187, 260, 289
    power spectrum, 186
    variance, 172
  perturbations, 71, 147, 149, 170, 176, 179, 288, 289
  polarization, 272, 274, 373, 385
  Ricci, 58, 59, 141, 149, 150
Tensor-to-scalar ratio, 187
Theory
  excursion-set, 353
  extended Press–Schechter, 353
Thermal
  neutrinos, 83
Thomson scattering, 113
Tight coupling, 119
Tightly-coupled
  solutions, 242
Tracers, 295
Transfer function, 217, 226
Two Degree Field Galaxy Redshift Survey (2dF), 10
Two-point functions, 412

**U**

Unbiased tracer velocities, 311
Universe
  background, 14, 34, 39, 58, 61, 138, 196, 297, 352
  Euclidean, 2, 4, 15, 26, 49, 51, 59, 79, 81, 136, 261–263
  inhomogeneous, 71, 78, 158, 378
  perturbed, 71, 73, 77, 79, 80
  structure, 10

**V**

Vacuum expectation value, 169
Vacuum states, 169
Value
  expectation, 422
Variance
  cosmic, 253, 262, 265, 288, 291, 407, 409, 417, 422
  perturbations, 171
  sample, 407
Vector
  contravariant, 25
  covariant, 25
  headless, 271
  perturbations, 147, 156
Velocity
  baryons, 132, 317
  dispersion, 82, 123, 330, 357
  distribution, 343
  divergence, 333, 337
  electrons, 119, 129
  equation, 129, 208, 215, 240
  fields, 337
  fluid, 123
  gas, 320
  irrotational, 121
  longitudinal, 121
  peculiar, 76, 297, 301, 303, 308, 309
  perturbations, 174
  relative, 223, 224

**W**

Wavevector, 120, 121, 130, 147, 152, 271, 274, 309
Weak lensing, 315, 320, 358, 366, 376, 389–391, 396, 407
  statistics, 388
Weakly Interacting Massive Particle (WIMP), 99
Weighting
  inverse-variance, 404
Weighting factor, 292
Wick's theorem, 339

**Z**

Zel'dovich approximation, 370