

S385

Cosmology and the distant Universe

Cosmology Part 1



This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2023. Second edition 2024.

Copyright © 2023, 2024 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd, 1 St. Katharine's Way, London, E1W 1UN (website www.cla.co.uk).

Open University materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using L^AT_EX.

Printed and bound in the United Kingdom by Halstan Printing Group, Amersham

ISBN 978 1 4730 3987 2

2.1

Contents

Introduction	1
Chapter 1 Introduction to cosmology and the expanding Universe	3
1.1 Key ideas in cosmology	3
1.1.1 The cosmological principle	3
1.1.2 The expanding Universe	5
1.1.3 The brightness of the night sky	9
1.2 The contents of the Universe	12
1.2.1 Particles and interactions	12
1.2.2 Matter and energy	14
1.2.3 Properties of gases	15
1.2.4 Interaction of matter and radiation	17
1.3 The big bang model	20
1.3.1 A brief history of the Universe	21
1.3.2 The cosmic microwave background	23
1.4 Summary of Chapter 1	25
Chapter 2 Tools for mapping space and time	27
2.1 Understanding space and time	28
2.1.1 Reference frames and relativity	28
2.1.2 Transformations between reference frames	31
2.1.3 Consequences of special relativity	33
2.2 Spacetime and metrics	36
2.2.1 Spacetime diagrams	36
2.2.2 Causality and simultaneity	40
2.2.3 Metrics in space and spacetime	42
2.3 Curved space and spacetime	45
2.3.1 Flat and curved geometries	45
2.3.2 Defining and measuring curvature	50
2.3.3 Curved spacetime and geodesics	54
2.3.4 Proper time	55
2.4 Summary of Chapter 2	57

Chapter 3 The geometry of the Universe 59

3.1 Gravity as geometry	59
3.1.1 Free fall and the equivalence principle	59
3.1.2 Einstein's field equations	62
3.1.3 Evidence for general relativity	64
3.2 Spacetime near planets and black holes	66
3.2.1 The Schwarzschild metric and its properties	67
3.2.2 Black holes and the event horizon	69
3.2.3 Behaviour of spacetime near black holes	71
3.3 The geometry of the expanding Universe	75
3.3.1 Cosmic time and co-moving coordinates	76
3.3.2 The Robertson–Walker metric	79
3.3.3 The Hubble parameter and the scale factor	80
3.4 Summary of Chapter 3	84

Chapter 4 Cosmological models and their key parameters 87

4.1 The Friedmann equations	87
4.1.1 The Friedmann equation	87
4.1.2 The fluid equation	92
4.1.3 The acceleration equation	94
4.2 Modelling the contents of the Universe	95
4.2.1 The equation of state	95
4.2.2 Models for matter	96
4.2.3 Models for radiation	99
4.2.4 Mixture models	101
4.3 Density and curvature	103
4.3.1 The critical density	103
4.3.2 Density parameters	105
4.4 The cosmological constant	106
4.4.1 Introducing Λ	106
4.4.2 Implications of the cosmological constant	108
4.4.3 Models of the Universe with a cosmological constant	109
4.5 Summary of Chapter 4	114

Chapter 5 Measuring cosmological parameters	117
5.1 Defining distance	117
5.1.1 Proper distance	118
5.1.2 Luminosity distance	122
5.1.3 Angular diameter distance	126
5.2 Measuring distances	129
5.2.1 Stellar parallax	129
5.2.2 Standard candles	133
5.3 Measuring H_0 and density parameters	138
5.3.1 Measuring the Hubble constant	139
5.3.2 Cosmological distance ladders	141
5.3.3 Measuring the density parameters for matter and Λ	144
5.4 Summary of Chapter 5	152
 Solutions to exercises	 155
 References and acknowledgements	 163
 Index	 167

Introduction

In this book you will study the fundamentals of cosmology, and how cosmological models are defined and used to describe the geometry and evolution of the Universe on the largest scales. There are five chapters in *Cosmology Part 1*, which develop the concepts and the mathematical tools of cosmological theory, and examine some of the observational evidence used to test it.

- Chapter 1 provides a general introduction to the science of cosmology.
- Chapter 2 develops the concepts, language and mathematics of special relativity and of curved geometries, which underpin cosmological theory.
- Chapter 3 explores the key ideas of general relativity and how they are used to build a mathematical description of the Universe.
- Chapter 4 introduces and examines the Friedmann equations, the basis of modern cosmological models, and defines the key cosmological parameters that can be used to test cosmological theory.
- Chapter 5 considers how measurements of distances to astronomical objects are used to measure cosmological parameters, and therefore test the theories introduced in previous chapters.

The next book you will study – *Cosmology Part 2* – continues by examining another main observational tool for cosmology, the cosmic microwave background (or CMB) radiation. It then provides an in-depth history of the Universe, from a ‘soup’ of energetic particles to the formation of the first stars and galaxies.

The exercises in each chapter are an important element of your learning. They are there to help develop and reinforce key ideas, and several important concepts in this book are developed through the exercises and nowhere else. Full solutions for the exercises are provided at the end of this book, but do try to complete them yourself before looking at the answers.

A table of physical constants is also given at the end of the book for use in your calculations. Be aware that, in some cases, your numerical answers to calculations may differ slightly from those in exercise solutions, depending on the precision used for constants and any intermediate rounding in the solution. Definitions for terms highlighted in **bold** may be found in the module glossary.

Throughout the text, coloured boxes are used to highlight particular types of information. Orange boxes highlight the most important equations and other key information. Turquoise boxes indicate additional information, such as reminders of concepts that you may have met in previous study, or ideas that are partly beyond the scope of the module but provide additional context. Blue boxes indicate where further, optional resources are available on the module website.

Finally, it is important to acknowledge some important influences on the writing of the S385 books. Some of the content presented here

builds on books written for the previous Open University module S383 *The relativistic universe*, namely Robert Lambourne's *Relativity, gravitation and cosmology*, Stephen Serjeant's *Observational cosmology*, and Ulrich Kolb's *Extreme environment astrophysics* (all published in 2010). We are grateful to the module team who produced those texts.

Many other excellent textbooks on cosmology and extragalactic astrophysics have influenced the development of the books for S385. We would like in particular to acknowledge Barbara Ryden's *Introduction to cosmology* (2017) and Andrew Liddle's *An introduction to modern cosmology* (1999). We recommend these books as optional further reading, but please note that they may differ from S385 in their choice of notation, convention, and the mathematical approaches used to set out some key equations.

Chapter 1 Introduction to cosmology and the expanding Universe

Cosmology is the scientific study of the Universe as a whole. It involves taking a ‘big picture’ view of the contents, geometry, history and potential future of the Universe.

Over the last century or so, advances in theoretical physics, telescopes and other astronomical technology, and most recently the power of modern computation, have combined to provide powerful ways to mathematically describe and understand the Universe and its evolution, and to test the predictions of cosmological theory with precise observational measurements. We don’t yet have all of the answers – there are some fundamental gaps in our knowledge – but modern cosmology provides a rich and powerful toolbox for trying to understand the Universe.

The main ingredients of modern cosmology are: a few basic assumptions about the nature of the Universe, Einstein’s theory of general relativity, which relates space, time, matter and energy, and the physics of radiation and matter and the ways they interact. This introductory chapter sets out the context of modern cosmological theory and some important astronomical observations that have motivated the development of those concepts, and continue to provide ways to test our theories.

Objectives

Working through this chapter will enable you to:

- describe the key principles that underpin modern cosmological theory
- apply basic concepts of observational astronomy and the physics of matter and radiation relevant to cosmology
- identify the main stages in the history of the Universe and place particular events onto its timeline
- summarise the primary evidence for the expansion of the Universe and the hot big bang model
- solve numerical problems relating to the contents and temperature evolution of the Universe, and the interaction of matter and radiation.

1.1 Key ideas in cosmology

1.1.1 The cosmological principle

Cosmology is built on some fundamental assumptions, which make it possible to construct a (comparatively straightforward) mathematical description of the geometry of the Universe and how it changes with time. A key assumption is that there is no privileged location in the Universe: if the properties of space (or spacetime, the nature of which we will

discuss later) are measured from any location in the Universe, then what we learn should be true of the Universe as a whole. This leads to the **cosmological principle**, an important starting point for all of cosmology.

The cosmological principle

On the largest scales, the Universe is assumed to be **isotropic** (i.e. the same in all directions) and **homogeneous** (i.e. the same in all locations).

You might immediately question this assumption. Our local environment in space does not look very homogeneous or isotropic: the Solar System contains one star and a variety of planets and other objects, all spread out in a very uneven way. But cosmology is the study of the Universe on the largest scales, and it is on much larger scales than the Solar System that we can test this principle.

Figure 1.1 shows a slice of the observed Universe on scales useful for constructing cosmological theories. You are seeing part of the **cosmic web**, a network of individual galaxies that form clusters and superclusters with hundreds to many thousands of galaxies in each. Superclusters span distances of hundreds of megaparsecs (see the box that follows shortly about distance units), and it is on these scales that the Universe is thought to be homogeneous.

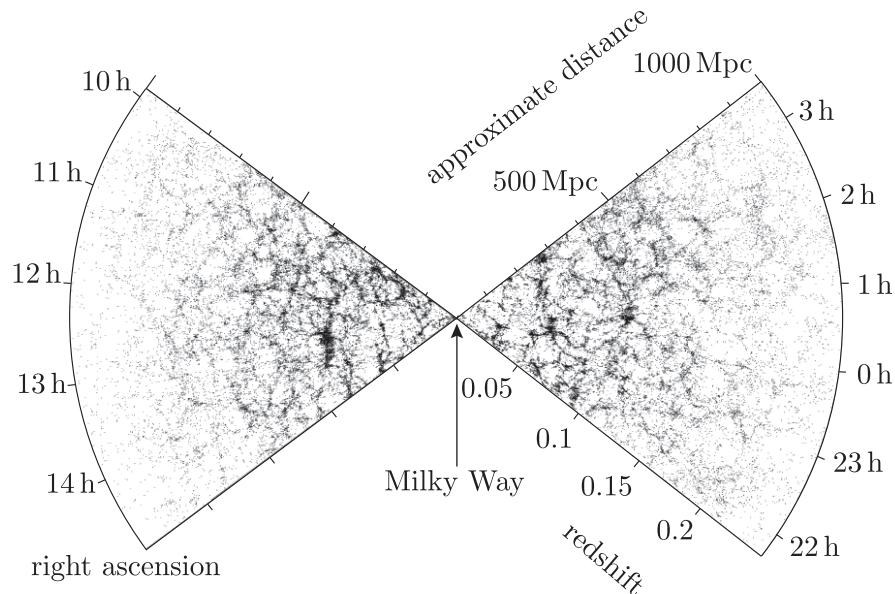


Figure 1.1 The large-scale structure of the Universe, mapped by the 2dF Galaxy Redshift Survey. Note that the apparent decrease in galaxy density at larger distances is due to it being harder to detect fainter objects, rather than there being fewer galaxies present.

Distance units in cosmology

The size scales considered in cosmology are vast, which means that ordinary SI units of metres and kilometres are impractical.

Astronomers usually express distances beyond our Solar System in units of parsecs (pc), where $1\text{ pc} = 3.086 \times 10^{16}\text{ m}$, or roughly 3.3 light-years. The distances between individual galaxies in the Universe are typically thousands to millions of parsecs, and so cosmological calculations often involve working with distances in units of megaparsecs (Mpc).

- Why does assuming the Universe is homogeneous make it easier to construct mathematical models that describe it?
- Homogeneity means that we can define universal parameters that describe the contents or behaviour of the Universe at a particular time (e.g. the overall density of matter), and that these parameters will have the same value everywhere. If the Universe were inhomogeneous then its key properties would depend on location as well as time, which would make cosmological models both more mathematically complex and harder to test with observations.

1.1.2 The expanding Universe

One of the most widely known cosmological facts is that the Universe is expanding. This idea has an interesting history. When Einstein's theory of general relativity was first applied to the geometry of the Universe in 1917, the solution had a peculiar and initially undesirable feature: that the Universe tended to contract or expand rather than remaining static.

At that time it wasn't yet proven that the Universe extended beyond our own galaxy, the Milky Way. Over the next few years the properties of galaxies beyond our own became firmly established; by 1930 it had been clearly demonstrated that nearby spiral galaxies were not part of our Galaxy, and were all moving away from us.

Figure 1.2 shows the observed relationship between the distance of galaxies and the speed at which they are receding, based on the work of Edwin Hubble and collaborators, which was published in 1929. Note that the vertical axis units are labelled incorrectly in this original plot – they are in fact km s^{-1} . Systematic errors in the 1920s distance measurements mean that the value of the slope of the relationship is also now known to be different.

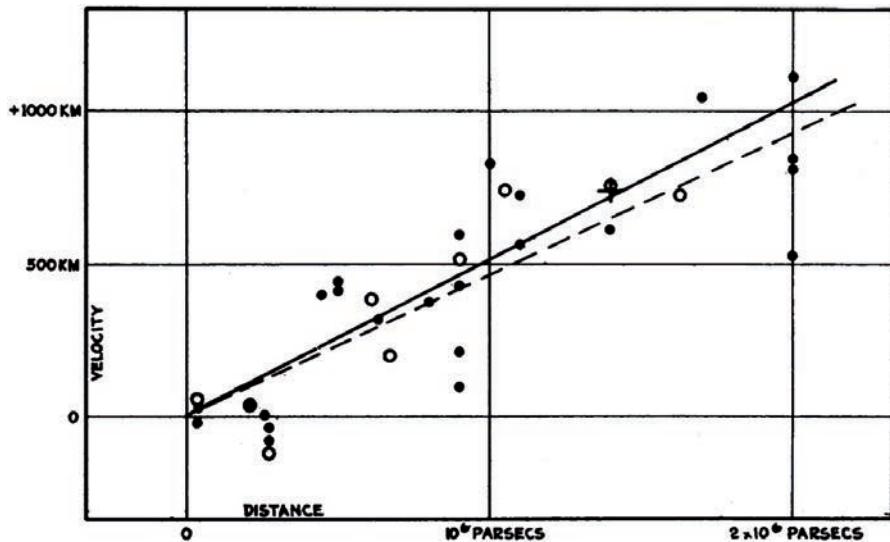


Figure 1.2 The famous result published by Hubble in 1929, showing a relationship between galaxy recession speed and distance. Such plots are now sometimes referred to as Hubble diagrams.

The galaxy recession speeds in Figure 1.2 were obtained by measuring redshifts of emission lines in the galaxies' spectra. A brief reminder of the concept of redshift is given in the following box.

Cosmological redshift

Redshift, z , measures the change in wavelength of radiation emitted from an astronomical body (such as a galaxy or quasar) caused by its movement away from the observer. It is defined as:

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} = \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} - 1 \quad (1.1)$$

where λ_{em} and λ_{obs} are the wavelengths of a spectral feature (e.g. an emission line, such as from atomic transitions) as emitted from the source and as measured by the observer, respectively.

Spectral shifts are caused by motion towards the observer (blueshifting) or away from the observer (redshifting). For distant galaxies, measured redshifts are dominated by the effect of cosmological expansion, which causes the apparent recession of galaxies, and so can be used as a distance measure.

As Figure 1.2 indicates, in the relatively nearby Universe there is a linear relationship between recession velocity and distance. This is known as the **Hubble–Lemaître law**, and can be written as

$$z = \frac{H_0 D}{c} \quad (1.2)$$

where D is the galaxy distance, typically measured in units of Mpc, c is the speed of light, typically measured in units of km s^{-1} , and H_0 is a quantity known as the **Hubble constant**.

Figure 1.3 shows a modern Hubble diagram, using redshift and distance measurements of supernova explosions in distant galaxies from several surveys. Note that the axes are swapped relative to Figure 1.2, with distance now on the vertical axis.

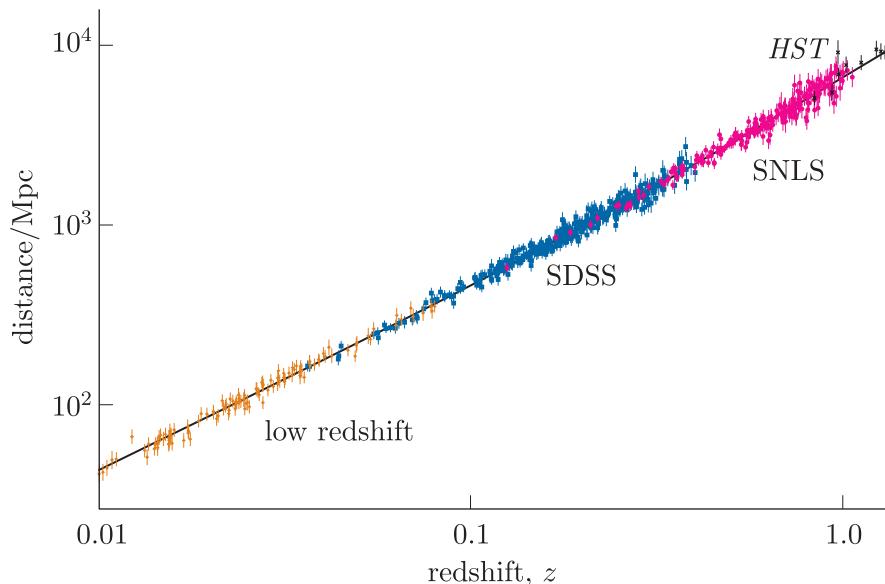


Figure 1.3 The relationship between distance and redshift from a compilation of supernova measurements (Betoule *et al.*, 2014), including a low-redshift sample, and samples from the Sloan Digital Sky Survey (SDSS), the Supernova Legacy Survey (SNLS) and *Hubble Space Telescope* (*HST*) observations.

If you look carefully at Figure 1.3 you will see that towards larger redshifts the relation begins to deviate from the straight line of the Hubble–Lemaître law. We will return to this observation later in the module.

- Based on Equation 1.2 and Figure 1.3, what is H_0 a measure of, and what units would you expect it to have?
- The Hubble constant is a measure of the rate of expansion of the Universe. Because redshift is a dimensionless quantity then, without applying any unit conversions, H_0 must have units of $\text{km s}^{-1} \text{Mpc}^{-1}$ to cancel out the units of D and c in Equation 1.2.

The units of the Hubble constant are a little unusual – kilometres per second makes sense for a rate of expansion speed, but what does the ‘per megaparsec’ part tell us? The answer lies in the relationship shown by Figure 1.2: we observe distant galaxies to recede more rapidly than nearer ones. In other words, the amount by which a region of space expands in a given time interval depends on the scale being considered. H_0 therefore measures the rate of expansion *over a fixed distance*: observers at any location at the current time will measure that, over a distance of 1 Mpc, the Universe expands at a rate of around 70 km s^{-1} .

What does it really mean for the entire Universe to be expanding? For this to be a well-defined concept, which can be rigorously tested with observations, it is necessary to set out a mathematical description of space and time.

The geometry of the Universe can be described by mapping out a four-dimensional geometry, with the usual three spatial dimensions plus the important dimension of time. This geometry is known as **spacetime**, and describes how the separation distance between locations in the Universe can be measured consistently by observers, irrespective of location. It is the expansion of this ‘fabric’ of spacetime that leads to the observation of the recession of distant galaxies. The concept of spacetime, and the mathematical framework to describe it, are not straightforward to grasp, and will be expanded on in the two chapters that follow.

In introducing the expansion of the Universe it is important to note that the term ‘Hubble constant’ is a misnomer, although it is widely used. Observations of the distant Universe demonstrate that the rate of expansion of the Universe is not constant with time. To avoid confusion, the subscript ‘0’ in H_0 is used to indicate that this is the expansion rate that applies at the *current* time, i.e. at the point in the Universe’s history at which we are measuring. When we treat the expansion of the Universe quantitatively in later chapters, you will instead be working with the Hubble parameter, which is the time-dependent generalisation of H_0 (the latter is therefore often referred to as ‘H-nought’, rather than as the Hubble constant).

The relationship described by the Hubble–Lemaître law allows cosmologists to draw conclusions about how the expansion of the Universe has changed with time over its history. Distance and time are intrinsically connected when studying the extragalactic Universe for a simple reason: the finite speed of light means that the light from distant galaxies has been travelling to us for billions of years, and so we are viewing them as they looked at a time when the Universe was very much younger. This link between distance and the time elapsed since light was emitted means that – because of the relationship shown in Figure 1.3 – astronomers can use redshift as a measure of cosmic time. High-redshift galaxies are effectively showing us the earlier history of the Universe, whereas low-redshift galaxies show us what is happening in the Universe closer to the present day.

- How does the cosmological principle help us to interpret any differences in the appearance of very distant galaxies compared to those nearby?
- The cosmological principle tells us that the galaxies in any region of the Universe at the present day should, on average, have the same properties. Any systematic differences in the appearance of distant galaxies compared to nearby ones must therefore be caused by the fact that we are observing them at different cosmic times, i.e. they must be due to how the Universe has evolved with time.

1.1.3 The brightness of the night sky

The expansion of the Universe, as demonstrated by redshift observations, is one of the foundational ideas of modern cosmology. Another direct form of evidence that we do not live in an infinite, unchanging (non-expanding) Universe dates back to a conundrum that was first identified in the early nineteenth century.

One of the most basic observations about the Universe is that the night sky is mainly dark. But should we expect that to be the case if the Universe is infinite and full of stars? In this section we will investigate what we would predict the night sky to look like in a static, infinite Universe. We will make use of some basic quantities in observational astronomy in the process, some of which are outlined in the following box.

Flux and luminosity

Flux is defined as the energy per unit time per unit area passing through a surface, for example the rate of visible-light photons from a star passing through a telescope detector at the Earth (typically measured in units of W m^{-2}). Flux, F , usually refers to the total energy across all relevant wavelengths. You may also meet the term **flux density**, which is flux per unit frequency or wavelength, measured in a narrower part of the spectrum.

Flux is related to **luminosity**, L , which is the total energy emitted per unit time by the system under consideration (for example, a star or a galaxy). The relationship between flux and luminosity is determined by geometry. If a distant object emits light isotropically, then the fraction of light that a detector of a particular size intercepts will depend on the distance from the object to the detector, leading to the following relation between flux and luminosity for an object at distance D from the Earth:

$$F = \frac{L}{4\pi D^2} \quad (1.3)$$

The **surface brightness**, Σ , of a square region of the sky with angular dimensions $\theta \times \theta$ is:

$$\Sigma = \frac{F}{\theta^2} \quad (1.4)$$

where F is the total flux of emission from that region of sky measured in units of W m^{-2} , and Σ will have units of $\text{W m}^{-2} \text{ deg}^{-2}$, assuming that the angle θ is measured in units of degrees.

To explore what we can learn about the Universe from the observation that the night sky is dark, we will first consider, in the example that follows, what we would predict to be the surface brightness of a typical star. Before that, the next box provides a brief reminder of some common notation associated with cosmology.

Common notation conventions

The symbol \odot is used in astronomy to indicate quantities associated with the Sun. Hence L_\odot is the Sun's luminosity, R_\odot is the solar radius, and so on. These quantities are also often used as units in astronomy, so that the luminosity of a galaxy, for example, can be expressed in units of solar luminosity (L_\odot , typeset without italic formatting), rather than in SI units of watts.

In comparison, a subscript asterisk can be used to denote the properties of another star, so M_* would signify a given star's mass.

Example 1.1

Assuming that the luminosity and intrinsic radius of a typical star match those of the Sun (L_\odot and R_\odot), calculate the ratio between the surface brightness, as observed at the Earth, of a typical star and that of the Sun.

Solution

Under this assumption, we can use the relationship between F and L to compare the surface brightness of the Sun, Σ_\odot , and that of a star, Σ_* , if the star is at a distance D_* from the Earth.

The ratio of the surface brightness of a star to that of the Sun is given by

$$\frac{\Sigma_*}{\Sigma_\odot} = \frac{(F_*/\theta_*^2)}{(F_\odot/\theta_\odot^2)} = \frac{F_*\theta_\odot^2}{F_\odot\theta_*^2}$$

Substituting in for luminosity using Equation 1.3 gives

$$\frac{\Sigma_*}{\Sigma_\odot} = \frac{(L_*/4\pi D_*^2)\theta_\odot^2}{(L_\odot/4\pi D_\odot^2)\theta_*^2} = \frac{L_* 4\pi D_\odot^2 \theta_\odot^2}{L_\odot 4\pi D_*^2 \theta_*^2} = \frac{D_\odot^2 \theta_\odot^2}{D_*^2 \theta_*^2}$$

where the (equivalent) luminosities have cancelled out, so that the ratio depends on the product of the distance and angular size of the two stars.

The final step is to relate the stars' angular sizes to their distances. The simple geometric considerations shown in Figure 1.4 illustrate that the angular diameter of a star is related to its size and distance by $\theta = 2R/D$, where we have used the small-angle approximation $\tan(\theta/2) \approx \theta/2 = R/D$.

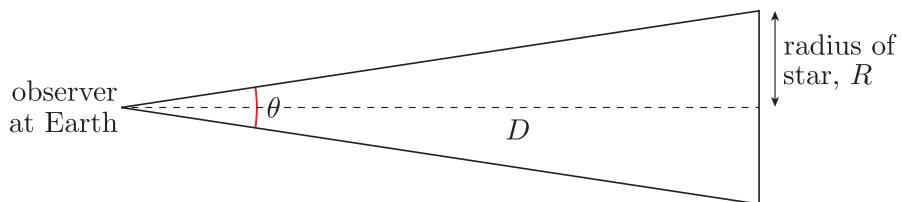


Figure 1.4 The relationship between distance D and physical radius R for an astronomical object spanning a small angle, θ , of the sky.

This means that

$$\frac{\Sigma_*}{\Sigma_\odot} = \frac{D_\odot^2(4R_\odot^2/D_\odot^2)}{D_*^2(4R_*^2/D_*^2)} = \frac{R_\odot^2}{R_*^2} = 1$$

where the last step follows because the star was assumed to be Sun-like (i.e. $R_* = R_\odot$). In other words, we conclude that the ratio between the surface brightness of a typical star and that of the Sun is 1.

We have just demonstrated that the surface brightness of a typical star of the same luminosity and radius as the Sun will be the same as that of the Sun. In other words, brightness per unit area of the sky is independent of distance for objects of the same intrinsic properties. Work through Exercise 1.1 to consider the implications of this conclusion further, by directly comparing the surface brightness of the Sun with the average surface brightness of the night sky.

Exercise 1.1

The Sun has an angular diameter of around 0.5 deg as seen from Earth, and a luminosity of $L_\odot = 3.8 \times 10^{26}$ W. Calculate its surface brightness, Σ_\odot , in units of $\text{W m}^{-2} \text{ deg}^{-2}$, and compare this value to the mean surface brightness of the night sky, which is $\sim 2 \times 10^{-13} \text{ W m}^{-2} \text{ deg}^{-2}$. (*Hint:* the Sun's distance from Earth is 1.5×10^{11} m.)

Of course, it is not surprising to discover that the surface brightness of the Sun is very much higher than that of the night sky – our own eyes tell us this! But if the Universe is infinite, and all the stars in the Universe have (roughly) the same surface brightness, shouldn't we see stars in *every* direction we look at night, and the sky be much brighter than we observe it to be?

This conundrum is known as **Olbers' paradox**, after Heinrich Olbers, who considered it in the 1820s. You may have questioned some assumptions that were made, but (for example) assuming that stars are less similar to the Sun than we have assumed does not solve the problem.

One possible solution to the paradox is that the Universe is not infinite in space. We cannot directly test this with astronomical observations, but we do now have considerable evidence that the Universe is not infinite in *time*. The Universe has a finite age, thought to be ~ 13.8 billion years, and the light from some stars has not yet had time to reach us.

- In a static (non-expanding) Universe with a finite age of ~ 13.7 billion years, how far away is the most distant star from which light has had time to reach us?
- Using the familiar relationship between speed (here c), distance (D) and time (t), the distance light could have travelled is $\sim 1.3 \times 10^{26}$ m, or ~ 4200 Mpc.

One of the main lines of evidence that the Universe is not infinitely old is that we observe it to be expanding, as discussed in the previous section. The expansion of the Universe also means that the most distant stars are receding further and further away. So Olbers' paradox is no longer a worry for modern astronomers, but it does provide clear evidence against the idea that we are living in a static, infinite Universe.

1.2 The contents of the Universe

Why does the Universe expand, and what controls its rate of expansion? The theory of general relativity provides a way of trying to answer these questions, and can be summed up in a famous quote from the physicist John Wheeler: ‘Space tells matter how to move; matter tells space how to curve’ (Misner, Thorne and Wheeler, 1973, p. 5). If we want to understand why and how space is expanding, it’s necessary to think about the matter contained within it. In fact, it is not just ‘matter’ that matters: the behaviour of spacetime is influenced by both matter and energy, which are linked by Einstein’s famous $E = mc^2$ equation.

Understanding the different forms of matter and energy present in today’s Universe, and how they interact with each other, also underpins our ability to make astronomical observations. Our main tools for measuring the Universe are electromagnetic radiation, particles such as cosmic rays and neutrinos, and, most recently, gravitational waves. All of these cosmic messengers are produced by the matter in stars and galaxies and, at the earliest times, by a ‘soup’ of particles filling the Universe. In the following sections you will review what we know about the contents of the Universe.

1.2.1 Particles and interactions

Modern particle physics is based on quantum field theory, which can successfully predict the properties and interactions of the families of particles that form the **Standard Model**. Most of the particles that are important for understanding the history of the Universe are part of the Standard Model.

Figure 1.5 shows the two major particle families: the **fermions**, which include three generations of matter particles, and the **bosons**, which carry force or mediate interactions. These families have different quantum mechanical properties. For example, fermions are subject to the **Pauli exclusion principle**, but bosons are not (a difference that is important for the structure of stars). In this module, the only boson that features heavily is the **photon**, the particle that mediates the electromagnetic force. Photons strongly influenced how the early Universe evolved, are crucial for the physics of stars and galaxies, and result in the images and spectra we measure with telescopes here at the Earth.

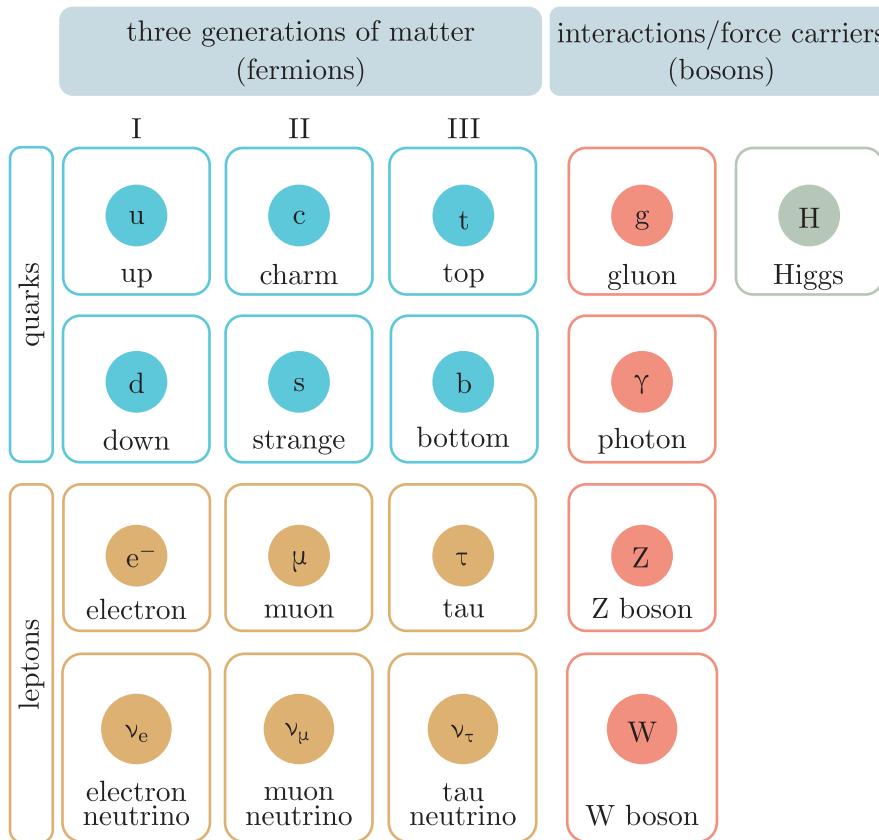


Figure 1.5 Particles in the Standard Model.

There are two families of fermions: the **quarks** and the **leptons**. These particles are distinguished by the types of fundamental interaction in which they participate. Leptons are subject to the **electromagnetic interaction** and the **weak interaction** (mediated by photons and the Z and W bosons respectively), while quarks also interact via the **strong interaction**, mediated by bosons known as gluons.

Considering the leptons first: although there are six types, as shown in Figure 1.5, in most cosmological contexts it is only necessary to consider the **electron** and its important ‘partner’, the electron **neutrino**. Electrons are, of course, an important constituent of atoms. There are also many situations in astrophysics where matter exists in the form of an ionised gas or **plasma**, in which ions and electrons are separated (and magnetic field effects can become more important than for neutral gases). Neutrinos are important by-products of a variety of weak interactions relevant to nuclear reactions occurring in the early Universe, as well as in stars.

Quarks also have six types, but as for the leptons it is only the first ‘generation’ that have much importance for the content of this module. Up and down quarks combine via the strong interaction to form the more familiar **proton** and **neutron**, each consisting of three quarks in a combination that yields electric charges of +1 and 0, respectively.

Each fermion also has an antiparticle equivalent, which is not shown in Figure 1.5. For example, the **positron** is the antiparticle counterpart of the electron, having the same mass but positive charge.

Protons and neutrons are types of **hadron**, a term that encompasses all particles made by assembling two or more quarks via the strong force.

Baryons are hadrons that contain an odd number of quarks, and so protons and neutrons are included in this category. Protons and neutrons make up most of the visible mass in the Universe, and so it is common in cosmology to refer to the **baryonic matter** of the Universe, meaning the ordinary matter that makes up stars, galaxies and interstellar and intergalactic gas. You will encounter this terminology regularly in discussions of cosmological models in later chapters.

Although most particles of interest to cosmology are part of the Standard Model, there is one important exception. ‘Invisible’ dark matter, not explained by the Standard Model, is thought to also be an important component of the Universe – a fact that can cause some scepticism about the whole endeavour of cosmology!

1.2.2 Matter and energy

There are several reasons to consider the overall contents of the Universe at this point. Figure 1.6 summarises the matter and energy contents of the present-day Universe and of the early Universe, according to current best observational estimates and cosmological theory. You will see later that the energy densities (the energy per unit volume) of different constituents of the Universe are important parameters in cosmological models, because Einstein’s field equations of general relativity tell us that the matter and energy content at any given epoch in the Universe’s history determines its rate of expansion at that time.

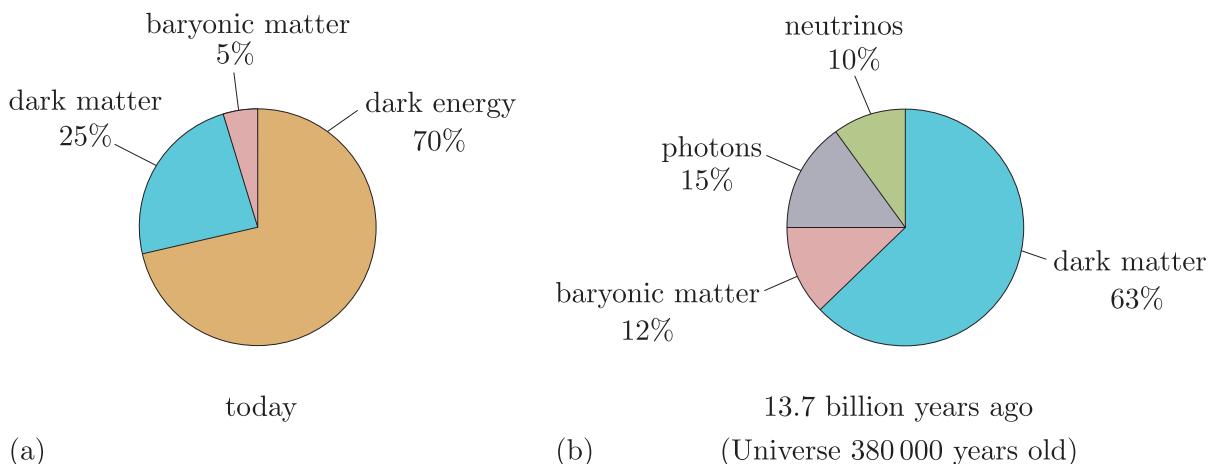


Figure 1.6 The matter and energy content of (a) the present-day Universe, and (b) the early Universe soon after the big bang.

Figure 1.6 also highlights the considerable limits of our current knowledge. All of the well-tested physics of the Standard Model discussed in the

previous section apply only to ~ 40 per cent of the matter and energy at early times, and an even smaller proportion of the overall matter and energy balance of the present-day Universe.

- What is the main difference between the make-up of the early Universe and that of the present day?
- The dominant energetic component of the present-day Universe is ‘dark energy’, but in the early Universe matter and radiation dominated, and a wider range of matter types contributed to its overall energy density.

Dark matter and **dark energy** are two of the biggest unknowns in cosmology. As mentioned in the previous section, it is currently thought that dark matter is a type of as-yet-undiscovered particle that is needed to explain the motions of stars in galaxies and the behaviour of galaxies and gas in clusters of galaxies, as well as several other important measurements. Dark energy is an as-yet-unexplained form of energy that was unimportant in the early Universe, but appears to be driving the rate at which the Universe is expanding in the present day. You will learn more about both of them in later chapters.

1.2.3 Properties of gases

The physics of matter and radiation in the Universe is primarily the physics of gases – or, more properly in many contexts, of plasmas. For most situations considered in this module this distinction can be ignored, and so for simplicity we will often use the term gas irrespective of whether or not the material is fully ionised.

One of the most important properties of gases for understanding the evolution of the Universe, and for many physical processes occurring in stars and galaxies, is density. Commonly, we will use n in this module to refer to particle number density (the number of particles per unit volume) and ρ to refer to mass density (the mass per unit volume).

The other fundamental property of gases to consider is temperature, T , which is a measure of the thermal or kinetic energy associated with the motions of individual particles of the gas. In a hotter gas the typical speeds of particles are higher than those in a gas at lower temperature. The populations of baryons and leptons within a gas will have a range of different speeds. When those speeds are sub-relativistic (i.e. well below the speed of light), interactions between the particles will result in a stable distribution of particle kinetic energies over time, or a situation of **thermal equilibrium**. Under these conditions the particles’ speeds, v , follow a **Maxwell–Boltzmann distribution**, as shown in Figure 1.7, and are described by

$$f(v) dv = \left(\frac{m}{2\pi k_B T} \right)^{3/2} 4\pi v^2 e^{-mv^2/(2k_B T)} dv \quad (1.5)$$

where $f(v) dv$ is the fraction of particles (of mass m) whose speed (magnitude of velocity vector) has a value between v and $v + dv$, and k_B is the Boltzmann constant.

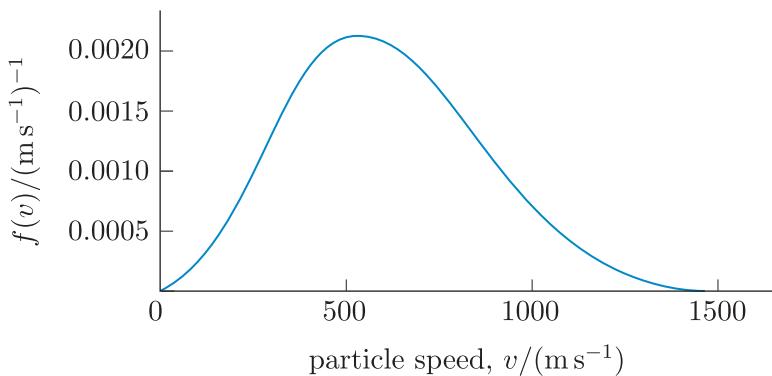


Figure 1.7 Example of a Maxwell–Boltzmann velocity distribution.

- Why is it very improbable for gas particles to have a speed at or close to zero?
- In a gas where most particles are moving at high speeds, it is very likely that any stationary particle will undergo a collision with a moving particle, which will cause the stationary particle to speed up.

It's important to be aware of this distribution of particle energies, because although the majority of interactions will involve particles with properties at the peak of the distribution, in some situations the presence of a 'tail' of higher-energy (faster-moving) particles is important. Equation 1.5 also explains why some other equations you will meet later in the module include an exponential factor that depends on particle kinetic energy (the mv^2 term) and temperature.

- Would you expect photons to follow a similar distribution? Briefly explain your reasoning.
- Photons always travel at the speed of light, so cannot have a distribution of speeds.

Although photons cannot have a distribution of speeds, those produced in a particular environment will have a distribution of *energies* (corresponding to the frequency or wavelength of light), and in situations of thermal equilibrium this can be used to assign an effective temperature to the radiation.

For matter particles with the most extreme energies, where their speeds approach the speed of light, particle velocity distributions no longer take a Maxwell–Boltzmann form; the somewhat different behaviour of such relativistic particles will be important in some situations considered in this module. Particle energies also depart from the Maxwell–Boltzmann distribution at high densities, where quantum mechanical effects, and the differing behaviours of fermions and bosons, become important.

One relationship that is different for (relativistic) gases made of photons rather than matter particles is that between pressure, P , and temperature, T . For a gas of particles in many astrophysical situations we can use the **ideal gas law**:

$$P = nk_B T \quad (1.6)$$

In contrast, a gas made of photons has the following relationship between pressure and temperature, with no density dependence:

$$P = \frac{1}{3}aT^4 \quad (1.7)$$

where a is the radiation constant.

1.2.4 Interaction of matter and radiation

Electromagnetic radiation was present in very large quantities in the early Universe, but the light that is observed at the Earth – and that allows us to study cosmology – has all been produced or affected by interactions involving baryons, electrons and/or ions. (Note that in this module we will use ‘light’ to refer to electromagnetic radiation across its whole spectrum, and will refer to ‘optical’ or ‘visible’ light at particular points if it is necessary to be more specific.)

- What are some of the processes by which light is produced in astrophysical environments?
- Light can be produced by particle–antiparticle annihilation and by any process in which particles undergo acceleration, including electromagnetic interactions between ionised particles in a gas. In an atomic gas, light of particular wavelengths is produced by atomic transitions in which electrons transition to a lower energy state, releasing photons in the process.
- What gas properties influence the production of radiation?
- Typically, gas density and temperature are the controlling parameters. Properties such as magnetic field strength can also be relevant.

Propagation of radiation in a gas

As well as determining the properties of the radiation we see, gas density and temperature also determine in what situations radiation escapes from its environment to travel to our telescopes, and thus what information the radiation is carrying about that environment.

A key concept here is **opacity** – the extent to which gas properties allow photons to travel through the medium without being absorbed or scattered by gas particles. If we know which absorption or scattering process is the dominant type of interaction in a particular environment, then it is possible to calculate the **mean free path**, which is the typical distance a photon can travel before interacting with a gas particle. For example, in a fully ionised gas – where ions and electrons are not combined into atoms – the most likely type of interaction for a photon under a wide range of conditions is **Thomson scattering**. Here, a photon’s energy and direction are altered by interaction with an electron.

The mean free path, λ , for electron scattering in a gas is

$$\lambda = \frac{1}{n_e \sigma_T} \quad (1.8)$$

where n_e is the electron number density and σ_T is the Thomson cross-section ($6.652 \times 10^{-29} \text{ m}^2$), which is analogous to a cross-sectional ‘target area’ for the photon to hit.

Exercise 1.2

The present-day mean electron number density in the halo (outer regions) of the Milky Way is $\sim 100 \text{ m}^{-3}$, while at a particular epoch in the early Universe it was $\sim 5.0 \times 10^9 \text{ m}^{-3}$. Calculate the mean free path of a photon in each of these environments in units of kiloparsecs, assuming electron scattering is the dominant interaction process for photons. Comment on how these distances compare to the size of a typical present-day galaxy (e.g. the Milky Way has a diameter of $\sim 27 \text{ kpc}$).

In the early Universe, photons could only travel on scales much smaller than individual galaxies before scattering off an electron, whereas photons produced in the halos of galaxies today can easily travel much larger distances without such interactions. This example highlights the importance of interactions between photons and gas particles for how light travels in the Universe, which is primarily controlled by gas density.

Ionisation of gas

Another very important interaction between radiation and matter concerns the **ionisation** of gas, which is dependent on gas temperature. The level of ionisation can affect the opacity of a gas, as well as determine which types of particle are available for other interactions.

The ionisation state of a gas of hydrogen in thermal equilibrium is described by the **Saha equation**. This equation compares the ratio of the number density of hydrogen atoms (n_H) to those of dissociated protons and electrons (n_p and n_e , respectively):

$$\frac{n_H}{n_p n_e} = \left(\frac{m_e k_B T}{2\pi\hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_B T}\right) \quad (1.9)$$

Here, m_e is the electron mass, k_B the Boltzmann constant, T the gas temperature, \hbar (h-bar, known as the reduced Planck constant) is equal to $h/2\pi$ where h is the Planck constant, and Q is the binding energy of a hydrogen atom, i.e. the energy an absorbed photon needs to have in order to ionise an atom of hydrogen ($13.6 \text{ eV} = 2.18 \times 10^{-18} \text{ J}$).

The ionisation fraction of a gas is defined as $X = n_p/n_b$, which is the ratio of the number density of free protons to that of all baryons (which, for a pure hydrogen gas is given by $n_b = n_p + n_H$). The next example explores this relationship.

Example 1.2

Show that for a pure hydrogen gas, in which $n_e \approx n_p$, the Saha equation can be rewritten in terms of the ionisation fraction as

$$\frac{1-X}{X} = n_p \left(\frac{m_e k_B T}{2\pi\hbar^2} \right)^{-3/2} \exp \left(\frac{Q}{k_B T} \right) \quad (1.10)$$

Solution

Examining the Saha equation (Equation 1.9), we can see that the right-hand side of the equation is completely determined by temperature, because all other terms are constants. We can therefore simplify the Saha equation temporarily by replacing the right-hand side with a constant depending only on T , which we'll call $C(T)$. Using the suggested assumption that $n_e \approx n_p$, Equation 1.9 becomes:

$$\frac{n_H}{n_p^2} = C(T)$$

The final expression we are aiming to derive involves only X on the left-hand side, while the right-hand side is equivalent to $n_p C(T)$. We can therefore rearrange our already simplified Saha equation to produce this same right-hand side:

$$\frac{n_H}{n_p} = n_p C(T)$$

We now need to express the left-hand side in terms of X . We start by expanding the expression for X in terms of the quantities in the Saha equation, n_H and n_p :

$$X = \frac{n_p}{n_b} = \frac{n_p}{n_p + n_H}$$

There are various ways we could manipulate this to find a relation between n_H/n_p and X . One is to take the reciprocal of both sides of the expression and then rearrange:

$$\frac{1}{X} = \frac{n_p + n_H}{n_p} = 1 + \frac{n_H}{n_p}$$

Hence

$$\frac{n_H}{n_p} = \frac{1}{X} - 1$$

and multiplying both sides by X and rearranging again gives

$$\frac{n_H}{n_p} = \frac{1-X}{X}$$

which can be substituted into the rearranged and simplified Saha equation above to give the required expression:

$$\frac{1-X}{X} = n_p C(T) = n_p \left(\frac{m_e k_B T}{2\pi\hbar^2} \right)^{-3/2} \exp \left(\frac{Q}{k_B T} \right)$$

Now try the following exercise to investigate ionisation fractions in different astrophysical situations.

Exercise 1.3

Calculate the ionisation fraction X for gas conditions corresponding to:

- (a) the Sun's outer regions (corona), where $T \approx 10^6$ K and $n_p = 10^{14}$ m $^{-3}$
- (b) a temperature of $T = 4000$ K and $n_p = 5 \times 10^9$ m $^{-3}$
- (c) a temperature of $T = 3000$ K and $n_p = 5 \times 10^9$ m $^{-3}$.

In all cases you may make the simplifying assumption that the gas is composed of pure hydrogen.

The preceding exercise shows that the ionisation fraction is very sensitive to the gas temperature, because of the exponential term on the right-hand side of the Saha equation. More generally, in this section you have seen that gas density and temperature have a strong influence on how radiation and matter behave. The next section sets out a broader overview of how the changing density and temperature of the early Universe have driven the evolution of matter, leading to the eventual formation of stars and galaxies.

1.3 The big bang model

The inevitable consequence of the Universe's expansion is that any region of space must have been smaller at earlier times than at present, so that matter and energy must have been more concentrated. The structures that we see in the present-day Universe – stars and galaxies and so on – must have formed from material that was previously concentrated into a very small volume. This scenario is known as the **big bang model**, in which everything that we observe today evolved from an initial state that was much smaller, hotter and denser than today's Universe.

Imagining an early Universe so different from our own environment is hard, and it is perhaps also hard to conceive that we could have got from there to here. But there is a wealth of evidence from different types of astronomical observations in support of the big bang model. This evidence includes a powerful, direct way of measuring the properties of the early Universe (e.g. its temperature and density): the cosmic microwave background (CMB) radiation.

In the remainder of this chapter we will sketch out the timeline of the history of the Universe to give you a ‘big picture’ outline, which you will be able to supplement with additional knowledge as you work through the rest of the module. You will also be introduced to the basic properties of the CMB and why it provides such a powerful tool for understanding the history of the Universe.

1.3.1 A brief history of the Universe

The big bang model does not attempt to explain *how* or *why* the Universe came into being. It only seeks to construct a history of how the Universe has evolved by extrapolating details from what we can observe (including the laws of physics operating in the present-day Universe). This story must be self-consistent: the laws of physics should not need to change, and the conditions in the early Universe, however extreme, must evolve naturally to produce the stars, planets and galaxies we see today, over 13.5 billion years later.

Figure 1.8 sets out the key evolutionary changes between the big bang and the present day. A lot of important action occurred in the first half-million years after the big bang, as spacetime expanded and matter cooled. The seeds of today's large structures – galaxies and groups and clusters of galaxies – were present in the cosmic soup of particles at early times. Then, under the influence of gravity, these small variations in density grew over billions of years to form the stars and galaxies we see in the night sky now.

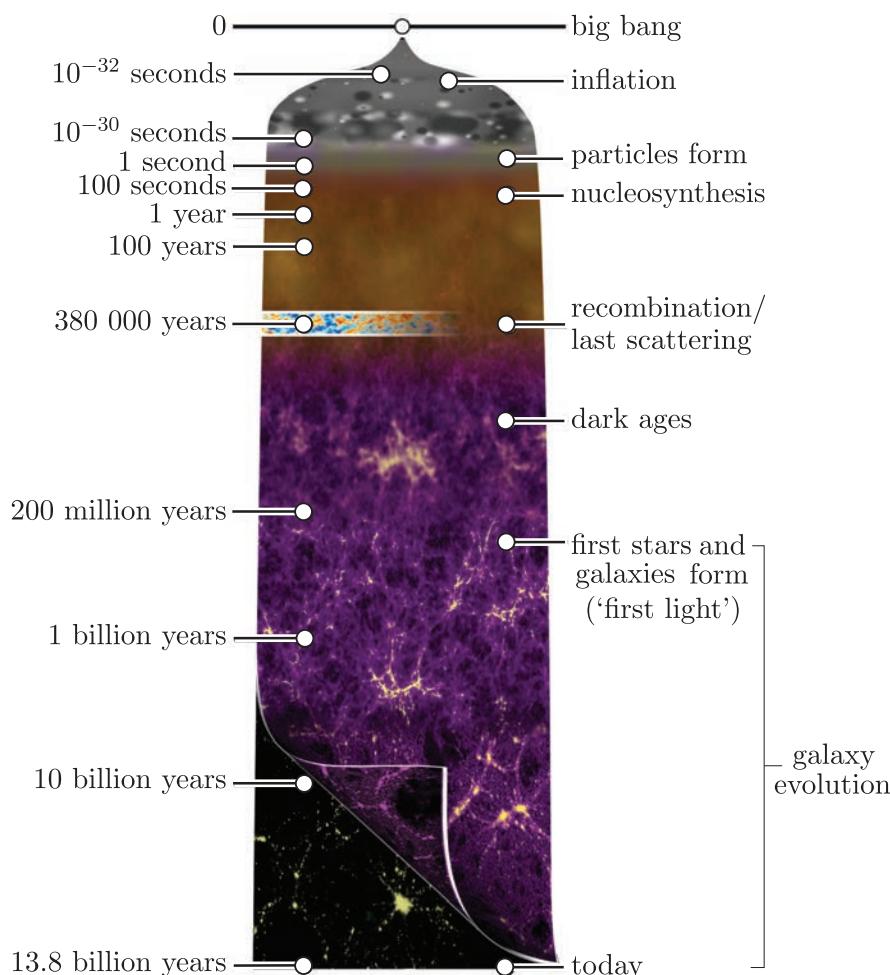


Figure 1.8 A simple timeline of the main stages in the history of the Universe. Note that this isn't a graph – the timeline is stretched to best enable all the important stages to be visualised.

A very brief description of the key periods in the timeline of the Universe is given below. As we move down the list the density of the Universe is decreasing, which reduces its energy per unit volume. Thus, temperature decreases with time too.

- **Inflation:** An almost incomprehensibly brief period, during which it is believed spacetime expanded by a very large factor, before slowing to a gentler expansion.
- **Particles form:** An initial cosmic soup of leptons, quarks, neutrinos and radiation cool, thereby allowing protons and neutrons to form (**baryogenesis**).
- **Nucleosynthesis:** When the Universe cooled sufficiently, it became possible for atomic nuclei to form and synthesise a small number of chemical elements (H, He, Li).
- **Recombination:** While high temperatures persist, gas remains ionised, because there are always many energetic photons to cause atoms to dissociate. Recombination describes the point at which the temperature became low enough to allow atoms to form.
- **Last scattering:** At high densities, ions, electrons and photons are constantly interacting, resulting in thermal equilibrium. By around the time of recombination, the density had dropped sufficiently for photons to escape from this gas, undergoing their ‘last scattering’ with the gas particles, after which the Universe became transparent to photons. The spherical ‘surface’ of this last scattering, as perceived from the Earth, delineates the earliest directly observable extent of the Universe, beyond which it was/is opaque.
- **Dark ages:** The early stages of **structure formation** – in which matter clumped together under gravity to form the seeds of stars and galaxies – are not visible to our telescopes, because the optical light produced in those environments was immediately absorbed by the gas.
- **First light:** The first stars and galaxies formed several hundred million years after the big bang, and their light ionised the surrounding gas. The Universe gradually became transparent to optical light.
- **Galaxy evolution:** Galaxies and their wider environments continued to evolve in complex ways over billions of years until the present day.

This timeline is the result of the combination of theoretical physics with testing via powerful observations. The concept of **lookback time**, an extremely useful consequence of the finite speed of light, is what allows such observational tests of cosmology. As we discussed earlier, the Universe is sufficiently large that we can observe objects and processes taking place at distances such that the light we measure left its originating location billions of years ago, when the Universe was a fraction of its current age. It is a remarkable property of the Universe that by looking as deeply as we can with our telescopes we are able to see how the Universe looked at an earlier point in time. Perhaps the most incredible such example is the subject of the next section.

1.3.2 The cosmic microwave background

Extraordinary claims require extraordinary evidence, and certainly some conclusions of cosmology are pretty extraordinary. But scientists have amassed rich and varied evidence in support of the currently favoured cosmological model. The **cosmic microwave background** (CMB) radiation is such a crucial part of this observational evidence that it deserves an introduction here. (In later chapters you will consider in detail the range of inferences that can be made from these observations.)

The CMB is a radio signal that pervades the Universe. It was first detected in 1965 by Penzias and Wilson, having previously been predicted as a consequence of theories of the early Universe. The history of its discovery is an interesting story, but one we don't have space to include here – it is well documented in science history books and online resources if you wish to learn more about it.

The two fundamental measurable properties of the CMB are its spectrum and its spatial distribution across the sky. Since its discovery, the CMB has been mapped in increasingly exquisite detail, most recently with all-sky maps obtained by two space missions: the *Wilkinson Microwave Anisotropy Probe* (WMAP) launched by NASA in the early 2000s and the European Space Agency's *Planck* mission a decade or so later.

Figure 1.9 shows the most detailed ever all-sky map of the CMB, made by *Planck*. The image uses a projection that maps the celestial sphere in a way that preserves the relative areas of different regions while showing the full sky. The colours indicate small shifts in the measured spectrum at different locations on the sky, which trace differences in the temperature (and hence density) of the material that produced the radiation. Structures are present both on large and small scales, and in a later chapter you will explore how these structures encode information about the early Universe.

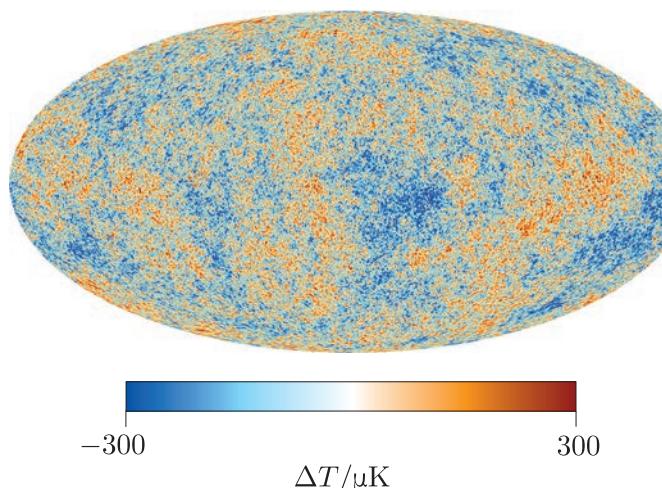


Figure 1.9 An all-sky map of the CMB radiation, as mapped by ESA's *Planck* mission. Colour indicates the deviation of temperature from the mean, ΔT , at each position on the sky, showing detailed structure that encodes information about the early Universe.

One of the most remarkable features of the CMB is how uniform it is when we measure it in all directions in space. The fluctuations shown in Figure 1.9 are variations of ~ 1 part in $>10\,000$, and so while the variations are very scientifically important (as you'll see later), the CMB has a uniform appearance on the sky in all directions when considered to an accuracy of much better than 1 per cent. This has profound implications, which we'll come back to in a moment.

A second important observation about the CMB is that it has a **black-body spectrum**: a characteristic dependence of brightness on wavelength, produced from regions that are opaque to radiation and in thermal equilibrium. The CMB's observed black-body spectrum reveals that the radiation comes from a region that is likely to be quite dense, so prior to escaping to travel towards us, the photons had a small mean free path and were constantly being scattered and/or absorbed and re-emitted.

Putting together these two features of the CMB – its high degree of uniformity and its black-body spectrum – leads to the conclusion that *the entire currently observable Universe* must have been in thermal equilibrium at the time the radiation was emitted. For this to be the case, the radiation must have been produced when regions of space were in close enough proximity to come into equilibrium. This makes it clear that the radiation is not being produced in the present-day Universe, when most of space is transparent to photons over cosmological distances, and the far reaches of the Universe are much too widely separated to come into equilibrium with each other.

Therefore, the CMB could only have been produced in a much denser early Universe. It is a key prediction of the big bang model that the photons produced when the Universe was last fully opaque should still exist, and so the CMB is one of three such predictions that are beautifully matched by observations. The other two predictions are the observed expansion of the Universe, and the abundances of key elements, both of which will be discussed in a later chapter.

Black-body radiation has the special property that the luminosity produced depends only on the temperature and surface area of the region from which it was emitted (i.e. unlike other processes, the amount of radiation produced doesn't depend on the quantity or density of matter present). The temperature, T , of a black body is related to the wavelength at which its brightness peaks. The observed wavelength peak of the CMB, at ~ 280 GHz, indicates a temperature of 2.7 K.

- If the CMB was produced in the very hot early Universe, then why is the temperature we measure so low (close to absolute zero)?
- The peak wavelength we measure here at the Earth doesn't tell us the temperature at which the radiation was produced. It is necessary to account for redshifting of the spectrum, i.e. the 'stretching' out of distances between wavefronts caused by the expansion of spacetime since the light was emitted.

Last scattering and the production of the CMB took place around 400 000 years after the big bang. According to the current cosmological model, the Universe has expanded by a very large factor since then, so that (although CMB photons were emitted throughout the entire volume of the Universe at the time) the CMB photons we observe here on Earth have travelled across a vast distance.

Exercise 1.4

The temperature of the CMB decreased as the Universe evolved according to $T \propto 1 + z$. If the ambient gas temperature at the time the CMB was emitted was $T \approx 3000$ K, calculate the redshift z corresponding to the time of CMB production.

The redshift of the CMB corresponds to a time when the Universe was around 0.4 million years old, or only around 3 per cent of its current age, so the CMB was, indeed, produced at a very early time in the Universe's history. Observations of the CMB therefore provide us with a powerful, direct way to measure properties of the early Universe.

1.4 Summary of Chapter 1

- The science of cosmology relies on the **cosmological principle**: the assumption that on the largest scales (hundreds of Mpc) the Universe is **homogeneous** and **isotropic**, i.e. it appears the same for all present-day observers at any location.
- The Universe is expanding, which is best described as an expansion of the spacetime intervals between fixed points. The Hubble parameter, H , measures the rate of expansion per unit distance. H changes with time as the Universe evolves, with its current value represented as H_0 , termed the **Hubble constant**.
- The main evidence for the expansion of the Universe comes from observing the **redshifts** of distant galaxies, which increase with distance, as captured in the **Hubble–Lemaître law**, demonstrating that all distant galaxies are receding.
- It is the finite speed of light that allows us to see how the Universe looked at an earlier point in time and study its evolution. The **lookback time** is that elapsed between the emission of light at its source and our detection of it here at the Earth.
- The present-day Universe contains a variety of forms of matter and energy, which have evolved over time, but its energy density is currently dominated by two components whose physics are not fully understood: **dark matter** and **dark energy**.

Chapter 1 Introduction to cosmology and the expanding Universe

- Many types of particle in the **Standard Model** are relevant to the science of cosmology. These include up and down **quarks** that interact via the **strong interaction** to form **protons** and **neutrons**, which are types of **baryon**. Baryons can combine via nucleosynthesis to form atomic nuclei, and with **electrons** to form atoms.
- Gas density and temperature influence the **ionisation** state of matter, the production of radiation, and the ability of radiation to escape from astrophysical environments (i.e. **opacity**).
- The observed expansion of the Universe motivates the **big bang model** in which the Universe began in a state of extremely high density and temperature, from which it has subsequently expanded in a process that continues to the present.
- Key periods in the history of the Universe include **inflation**, **baryogenesis**, **nucleosynthesis**, **recombination**, **last scattering**, the **dark ages**, **structure formation**, **first light** and **galaxy evolution**.
- The **cosmic microwave background** (CMB) radiation provides evidence for the hot big bang model.
- The CMB has a **black-body spectrum** and is spatially uniform, with low-level fluctuations that enable precision measurements of parameters that describe the Universe's subsequent evolution.

Chapter 2 Tools for mapping space and time

The science of cosmology involves attempting to map the Universe in space and to consider its evolution in time. You saw in the previous chapter that distance and time measurements in astronomy are closely linked: the finite speed of light means that distant parts of the Universe can only be observed as they appeared at times in the long-ago past. Our maps of the night sky are therefore maps of time as well as space, recording the history of the Universe.

In this chapter you will learn about a further way in which space and time are entwined. The sections that follow summarise some of the key concepts of special relativity and some of their non-intuitive consequences. The discussion of special relativity in this chapter forms the starting point for building the conceptual framework that underpins cosmological theory.

Online resources: introductory special relativity

If you have not previously studied Einstein's theory of special relativity, or would like a refresher, you may find the online resources for this chapter helpful.

Objectives

Working through this chapter will enable you to:

- explain the importance of reference frames for making measurements related to space and time intervals
- manipulate the Lorentz transformations to compare measurements in different frames and explore some non-intuitive consequences of special relativity
- define the concept of a metric, and apply it to examples of two- and three-dimensional geometries and four-dimensional spacetime
- explain how the curvature of geometric spaces can be measured, and describe the key differences between flat and curved spaces
- solve numerical problems relating to the curvature of two-dimensional surfaces.

2.1 Understanding space and time

2.1.1 Reference frames and relativity

If you have ever sat in a moving car watching roadside trees appear to zoom away from you, or in a train carriage looking at another train through the window and been unsure which train is stationary and which is moving, then you have everyday experience of the concept of **reference frames**. Any time we measure a speed it is in relation to a particular reference frame. You are unlikely to be reading this book from anywhere other than one particular rapidly rotating planet in a solar system moving at around 200 km s^{-1} around the centre of our Galaxy. However, the speed that you are moving relative to other parts of the Universe is usually irrelevant for considerations of how objects move in your local environment.

A basic concept in relativity is the **inertial frame**. An inertial frame is a reference frame that is not accelerating, in which the laws of motion (e.g. Newton's laws) apply.

Figure 2.1 shows two inertial frames, which are moving relative to each other at a constant speed, V . Each frame is defined by a set of coordinate axes: x , y and z in frame S, and x' , y' and z' in frame S'. It is usual in special relativity problems to define the x -axis as the direction of relative motion between the frames. Likewise, in the standard configuration for special relativity problems the frame origins are assumed to coincide at a particular time, $t = t' = 0$.

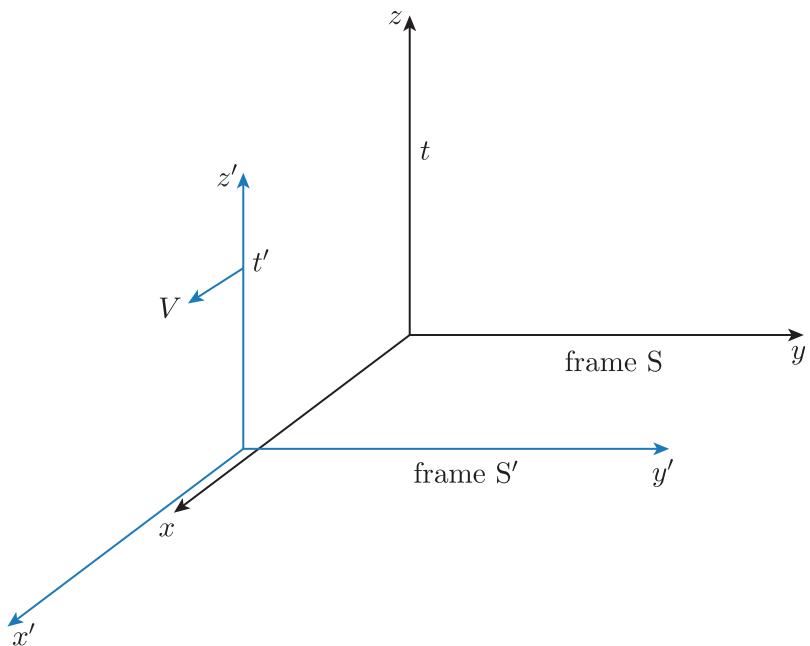


Figure 2.1 Two inertial references frames with coordinate axes labelled.

The special theory of relativity was born from the need to reconcile two well-evidenced statements ('postulates'):

The postulates of special relativity

- The laws of physics operate in the same way in all inertial frames.
- The speed of light in a vacuum has the same constant value, $c = 3 \times 10^8 \text{ m s}^{-1}$, in all inertial frames.

Both of these statements can be experimentally verified. The constant speed of light was first established via experiments in the late nineteenth century. A famous example is the Michelson–Morley experiment, which ruled out the presence of an 'ether': an all-pervading material, thought at the time to be a medium through which light waves travelled.

But the constancy of the speed of light is hard to reconcile with our intuitive understanding of how inertial reference frames behave. To understand why this is a problem, we need to introduce (or perhaps remind you of) some key definitions.

Special relativity terminology

- An **event** is an occurrence that takes place instantaneously (i.e. it does not extend for a significant length of time) at a fixed location in space.
- A **coordinate system** is a set of axes that define an inertial reference frame, e.g. x -, y - and z -axes, together with a time axis, t .
- An **observer** is someone who can make measurements of events, noting down the coordinates at which they occur within their own inertial frame.

In the following example you will consider the problem of how a moving object is measured in two frames, S and S' , where those frames are defined in the same way as in Figure 2.1.

Example 2.1

Figure 2.2 depicts a scenario at a train station, in which two observers experience an event from different reference frames.

- Observer B is on a train that is slowly passing through the station at a speed of $V = 20 \text{ km h}^{-1}$.
- Observer A is standing still on the station platform.

Observer B throws a ball down the train carriage at a speed of 10 km h^{-1} in the direction opposite to the train's direction of motion, i.e. in the $-x$ direction.

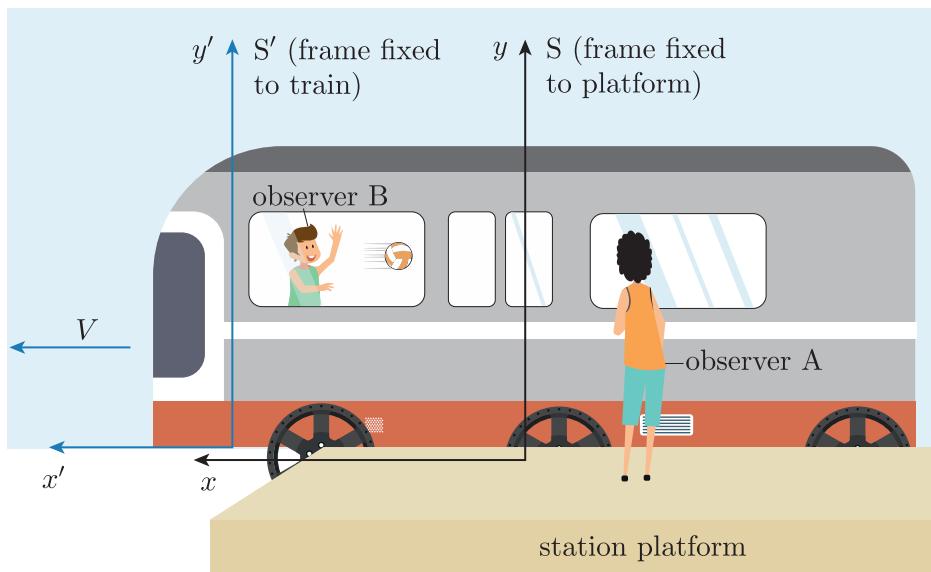


Figure 2.2 The reference frames of observers A (on a station platform) and B (on a train leaving the station at speed V in the x direction).

Taking S as the reference frame of observer A (on the platform), and S' as the frame of observer B (on the train), determine the speed of the ball as measured by each of the two observers. In other words, calculate and compare dx/dt and dx'/dt by considering how the coordinate position in each frame must change with time.

Solution

The ball's speed in B's frame, S' , is dx'/dt , where the prime ($'$) superscript indicates that we are referring to coordinates in frame S' . For A (whose frame is S), the ball's speed is given by its change of position relative to their coordinate x , so dx/dt .

The question tells us that in frame S' the ball's velocity dx'/dt is -10 km h^{-1} . To work out what A sees in frame S, we need to think about how the coordinates x and x' are related. If Δt is the time elapsed since $x = x'$ (i.e. the point at which the two sets of coordinates coincided), then in that time interval observer A will see a given location on the train changing position by $V\Delta t$. The x -coordinate location of the end of the train will be increasing, whereas its x' -coordinate will not, because the S' -coordinates are fixed to the train. Therefore, the x - and x' -coordinates are related by:

$$x = x' + V\Delta t$$

where $V = 20 \text{ km h}^{-1}$ is the relative speed of the two frames in the x direction.

Because we want to know the rates of change of x and x' , we can take the derivative of the relation above to get

$$\frac{dx}{dt} = \frac{dx'}{dt} + V$$

Therefore, we find that the velocity of the ball as observed by A is $-10 \text{ km h}^{-1} + 20 \text{ km h}^{-1} = 10 \text{ km h}^{-1}$. In other words, A will see the ball moving away from them, in the same direction that the train is moving, but it will be moving away more slowly than the train itself. So for A, the ball still appears to be moving down the carriage away from observer B, as expected on the basis of everyday experience.

The example showed that rates of change in the position of objects will be different when measured from different frames. But what if, instead of considering an object like a ball, we consider how a beam of torchlight produced by B travels down the train?

If we apply the same logic as we did in the example of the ball we should conclude that A observes the light beam to be travelling at a speed of $-c + 20 \text{ km h}^{-1}$, but this isn't consistent with the second postulate of special relativity. Both A and B *must* observe the light to travel at the same speed, so the logic we have applied in the example must have a gap in it. It is this gap that shows the need for a more sophisticated theory to describe the relationship between how different observers measure physical behaviour.

- In comparing measurements in two reference frames, what assumption was made in Example 2.1 about how the observers measure time intervals?
- We assumed that observers A and B measure time progressing (e.g. their clocks 'ticking') at the same rates.

In the next section you will see that the assumption that all observers measure time in the same way has to be sacrificed to enable the two postulates of special relativity to be reconciled in a logical way. The consequences of this are at odds with our everyday experience, but are well verified by a variety of experiments.

2.1.2 Transformations between reference frames

The example in the previous section used everyday intuition to consider how distance intervals and speeds are measured differently from two reference frames. This intuition can be summarised by writing a set of equations to transform coordinates between two inertial frames whose relative velocity in the direction of the x -coordinate axis is V (e.g. as shown in Figure 2.1):

$$x' = x - Vt$$

$$y' = y$$

$$z' = z$$

$$t' = t$$

These equations are the **Galilean transformations**. But, as you saw in the previous section, they run into trouble when we try to consider the behaviour of light. Instead, Einstein showed in 1905 that there is a set of transforms between reference frames that avoid this logical contradiction. These are known as the **Lorentz transformations** – after Hendrik Lorentz, who first determined them (though with a somewhat different interpretation to our modern understanding). They are summarised as follows:

The Lorentz transformations

$$x' = \gamma(x - Vt) \quad (2.1)$$

$$y' = y \quad (2.2)$$

$$z' = z \quad (2.3)$$

$$t' = \gamma(t - Vx/c^2) \quad (2.4)$$

where the standard assumption is made that the x -coordinate axis is aligned with the direction of relative motion. The quantity γ , known as the **Lorentz factor**, depends on the relative speed of motion V , and is defined as

$$\gamma = \frac{1}{\sqrt{1 - V^2/c^2}} \quad (2.5)$$

- What are the two key differences between the Galilean and Lorentz transformations?
- Although the relation between x and x' has a similar form for the two transformations, in the case of the Lorentz transformations it includes the additional factor of γ , which depends on speed. The other difference is that the time coordinate undergoes a transformation in the Lorentz case and not the Galilean one.

The following exercise will help you to explore further how these two sets of transformations relate to the world around us.

Exercise 2.1

Calculate the Lorentz factor when the relative speed is: (a) an everyday speed of 20 km h^{-1} ; (b) 90% of the speed of light.

*Optional Python extension:** you may find it interesting to write a short piece of Python code to plot how the Lorentz factor depends on speed, for values ranging from $0.1c$ to c .

*We are using the term ‘Python’ throughout this module to refer to the Python® programming language, as developed by the Python Software Foundation.

The exercise above shows that our intuitive Galilean view of relativity works reasonably well for speeds we experience in everyday life, because the Lorentz transformations reduce to the Galilean forms at low speeds. However, the large value of the Lorentz factor for objects travelling close to the speed of light leads to effects that disrupt our everyday expectations. In particular, you have now seen that measurements of time and spatial coordinates in different reference frames differ dramatically when the relative speed of the different frames becomes large.

2.1.3 Consequences of special relativity

By breaking our assumption that all observers must agree about the time intervals between events, the Lorentz transformations resolve the contradiction between the behaviour of ordinary objects (e.g. the train and the ball in Example 2.1) and that of light.

To better understand the consequences of the Lorentz transformations, it is necessary to consider how they apply to **intervals** of space and time. Consider two events, labelled 1 and 2, which occur in frame S at times t_1 and t_2 , and at two separate locations, x_1 and x_2 . The events are observed from two frames, S and S' , the relative speed of the latter being V in the x direction. In frame S, the time interval between the two events is $\Delta t = t_2 - t_1$, and the space interval, i.e. the difference between the locations of the two events, is $\Delta x = x_2 - x_1$.

We can use the Lorentz transformation equations to obtain expressions for $\Delta t'$ and $\Delta x'$, namely the intervals between the same two events as measured from frame S' . We can first write separate expressions for the coordinates of events 1 and 2 in S' (noting that there is no change in the y - and z -coordinates, so they can be omitted):

$$\begin{aligned}x'_1 &= \gamma(x_1 - Vt_1) \\x'_2 &= \gamma(x_2 - Vt_2) \\t'_1 &= \gamma(t_1 - Vx_1/c^2) \\t'_2 &= \gamma(t_2 - Vx_2/c^2)\end{aligned}$$

Defining $\Delta t' = t'_2 - t'_1$ and $\Delta x' = x'_2 - x'_1$, we can write expressions for these two intervals in frame S' by taking, in turn, the difference between the two x expressions and then the two t expressions:

$$\begin{aligned}\Delta x' &= x'_2 - x'_1 = \gamma(x_2 - Vt_2) - \gamma(x_1 - Vt_1) \\&\quad \Delta t' = \gamma(t_2 - Vx_2/c^2) - \gamma(t_1 - Vx_1/c^2)\end{aligned}$$

These expressions can be simplified to give two equations for the transformation of intervals, in a form similar to the original transforms:

$$\Delta x' = \gamma(\Delta x - V\Delta t) \tag{2.6}$$

$$\Delta t' = \gamma(\Delta t - V\Delta x/c^2) \tag{2.7}$$

Exercise 2.2

Show that Equations 2.6 and 2.7 can be rearranged to find similar expressions for Δx and Δt , each only in terms of coordinates in the S' frame. Comment on how the resulting expressions compare to those for $\Delta x'$ and $\Delta t'$.

Now that we have expressions for how intervals of time and space transform between reference frames, we can investigate one of the most interesting and well-known consequences of special relativity – the idea that time is relative, and ‘moving clocks run slow’. Example 2.2 explores this idea further.

Example 2.2

A short-lived particle is created in a particle physics experiment, and travels at very high speed relative to an observer in the lab. Its time and position of creation are t_1 and x_1 , respectively. The particle is then seen by the same observer to decay at a later time, t_2 , when it has moved (only in the x -direction) to a new location, x_2 . The particle’s lifetime, defined for a particle at rest, is known from theory to be 2.2×10^{-6} s.

Find an expression for the time interval between particle creation and decay, as measured by the lab observer, assuming the particle is travelling at a constant speed V in the x -direction.

Solution

It is always important in special relativity to first think carefully about defining reference frames. The measured lifetime corresponds to a time interval in a reference frame in which the particle is at *rest*, and so we define reference frame S' as a frame that moves with the particle.

Therefore, this reference frame moves at a constant speed of V with respect to the laboratory (which we define as the reference frame S of the observer).

The next step is to determine what information we have, and what information we need to work out. Because we only have two events to consider (particle creation and particle decay) there are only two intervals to evaluate in each reference frame, as summarised in Table 2.1.

Table 2.1 Information relating to particle observations in a lab.

Interval	Description	Expression	Value
Δx	distance travelled in S	$x_2 - x_1$	unknown
$\Delta x'$	distance travelled in S'	$x'_2 - x'_1$	0
Δt	particle lifetime in S	$t_2 - t_1$	requested quantity
$\Delta t'$	particle lifetime in S'	$t'_2 - t'_1$	$2.2 \mu\text{s}$

We need to find an expression for the particle lifetime observed in S, namely Δt . A crucial piece of information is that $\Delta x' = 0$: because the particle is not moving in its own frame, S', the locations of the particle's creation and decay must be *the same*. This means that we can determine Δt if we have an expression that involves any of the intervals apart from the unknown Δx .

We can use the interval transforms from the solution to Exercise 2.2 to show that

$$\Delta t = \gamma(\Delta t' + V\Delta x'/c^2)$$

and because $\Delta x' = 0$ in this scenario, we find that

$$\Delta t = \gamma\Delta t' \quad (2.8)$$

Because $\gamma > 1$ for all values of V (see Equation 2.5), the lab-based observer will always measure a longer lifetime for the particle than the lifetime of the particle at rest (i.e. that period measured in a reference frame in which the particle is stationary).

Example 2.2 provides an illustration of **time dilation**: for two events observed to occur at the same location in a reference frame S', an observer in reference frame S will measure a longer time interval than elapses in frame S'.

- If there are observers in both frames S and S', and each sees the *other* frame in relative motion, doesn't this lead to a logical contradiction in which each observer measures a longer time interval than the other?
- No – the situation is not symmetric: in one frame the two events occur at the same x -coordinate (e.g. in a frame moving with a relativistic particle, the particle's location does not change with time), while in the other frame the x -coordinate changes, meaning that setting up the reverse problem and applying the Lorentz transformations will give a different result.

It is important to recognise this asymmetry in how time dilation works: the dilation is measured by the observer who *does not* see the two events as being co-located.

Special relativistic effects such as time dilation (and a similar effect known as length contraction) seem odd and counter-intuitive. It is important to emphasise that these effects only become important for relative speeds that are a significant fraction of the speed of light. But there are many contexts – both in everyday life and in physics research – where they matter, and they have been repeatedly tested and confirmed. These contexts include particle physics experiments, global positioning system (GPS) technology, and a variety of well-studied astrophysical situations including jets from black holes and the explosions of massive stars.

Online resources: special relativity applications

We are only able to touch on the implications and applications of special relativity in this chapter. For more information and opportunities to put these ideas into practice, see the online resources for this chapter, which include further content on special relativity from the OU Stage 2 physics curriculum.

2.2 Spacetime and metrics

You have seen in the previous sections that space and time cannot be fully separated in physics – there is no *absolute* time that all observers can agree on, because it depends on your frame of reference. This idea led to the fundamental concept of spacetime, a four-dimensional union of space and time, introduced by Hermann Minkowski (a teacher of Einstein). The mathematical framework developed to describe spacetime makes it possible to set out physical laws that *can* be agreed on by all observers. In this section you will explore the geometry of spacetime and some of its implications for cosmology.

2.2.1 Spacetime diagrams

Some of the biggest challenges in cosmology come from trying to understand **causality** in the history of the Universe. How did the vast and varied Universe we observe today evolve from the very different conditions that observations show existed in the early Universe? As we noted in Chapter 1, a particularly interesting question arises from our observations of the CMB – how can it be so uniform when it originates from regions now vastly separated in space?

Minkowski introduced a helpful tool for visualising the possible connection (or lack of connection) between particular events. Figure 2.3 shows an example of a **spacetime diagram**, plotting one spatial dimension (x) against time. The vertical axis corresponds to time measured in an inertial reference frame S , and is plotted as ct . This approach gives the vertical axis the same units as distance x , plotted on the horizontal axis. A second inertial frame, S' is represented by the blue, diagonal axes labelled as ct' and x' .

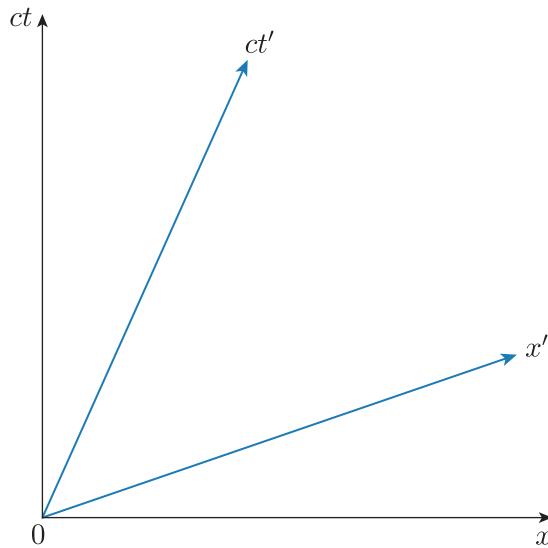


Figure 2.3 Example of a Minkowski spacetime diagram for an inertial reference frame S, with additional (blue) axes representing a second reference frame, S'.

Any fixed point on the diagram with coordinates (x, ct) corresponds to an event. But these coordinates refer only to that event in frame S. The next example explains the meaning of the diagonal axes shown in Figure 2.3.

Example 2.3

Consider an inertial frame S' with velocity V relative to frame S. The x' -axis of a spacetime diagram consists of all the events for which $ct' = 0$, and conversely the ct' -axis will be a line describing where $x' = 0$. (There is a matching relationship between the x - and ct -axes in relation to frame S.) Use the Lorentz transformations to find mathematical expressions for the straight lines corresponding to the x' - and ct' -axes on a plot of x vs ct .

Solution

Because the spacetime diagram is drawn with x, ct as its primary axes, we are first looking for the relationship between x and ct that holds true for all events on the x' -axis (and then similarly for the events on the ct' -axis).

Starting from the Lorentz transformations (multiplying the t' expression by the constant c) we can deduce the following:

$$\begin{aligned} ct' &= \gamma(ct - Vx/c) \\ x' &= \gamma(x - Vt) \end{aligned}$$

The question noted that the x' -axis is where $ct' = 0$. We can therefore set the left-hand side of the ct' transformation to zero to get an expression that must hold true for all events on the x' -axis:

$$0 = \gamma(ct - Vx/c)$$

Helpfully, this describes a relation between the S-coordinates of the events (x, ct) , which is what we are looking for. Rearranging to the familiar form for a straight line gives:

$$ct = (V/c)x$$

In other words, the x' -axis on the diagram is a straight line that goes through the origin and has slope V/c .

The ct' -axis can be found similarly, by setting $x' = 0$ in the second Lorentz transformation equation above (for x') and multiplying by c , to derive $ct = (c/V)x$.

The time and position axes of the S' frame therefore appear as diagonal lines with a slope that depends on the relative speed between the two frames, V .

- How does the appearance of the spacetime diagram alter if the relative speed of the frames, V , is increased?
- A higher value of V increases the slope of the x' -axis and decreases that of the ct' -axis, so that they move closer together. The equations for both of the S' -axes are shown in Figure 2.4.

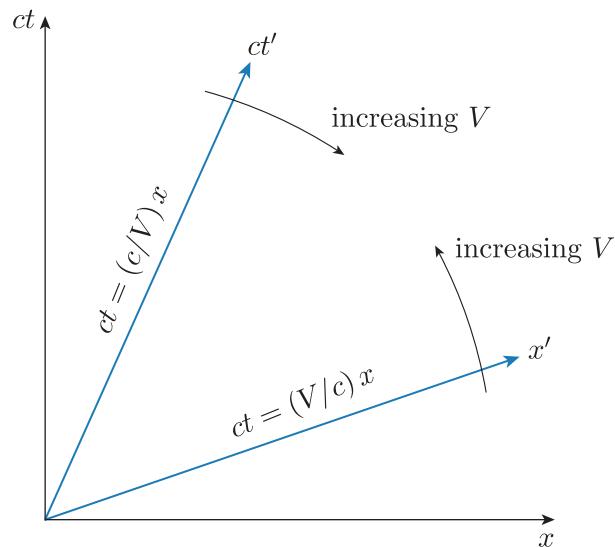


Figure 2.4 The effect on the S' -axes in a spacetime diagram of changing V .

Figure 2.5 shows how to read off the coordinates of an event from the diagram in both the S and S' frames. Three events are shown, two of which lie on the S' axes. Let's consider how to read off the coordinates of event 1 in each frame.

In frame S, event 1 has coordinates (x_1, ct_1) , but its coordinates in frame S' must be different. To read them off the diagram it is necessary to draw lines that intersect at event 1 and are parallel to each of the S' -axes; then the coordinates (x'_1, ct'_1) can be read from where these parallel lines intersect with the other primed axes, as shown in the diagram. The result

is that (coincidentally) event 1 has the same x' -coordinate as event 3, and the same ct' -coordinate as event 2.

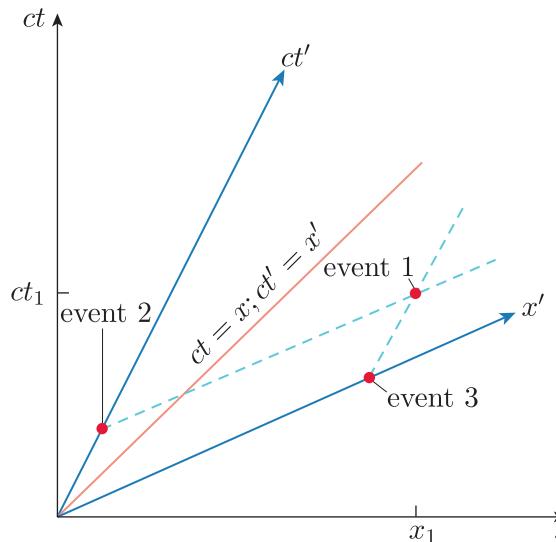


Figure 2.5 A spacetime diagram recording the coordinates of three events in the reference frames S and S' . A red line is also shown that plots $ct = x$ (which is equivalent to $ct' = x'$).

Another useful concept when using spacetime diagrams is the **world line**. A world line is the path that a particle (or another object; for example, a spacecraft) travels through spacetime. This too can be represented as a line on a spacetime diagram.

- What would the world line look like on a spacetime diagram for a particle measured in frame S' to be travelling at $0.9999c$?
- A particle travelling very close to the speed of light will travel along a line of slope ~ 1 on the diagram – in other words, its world line is very close to the red line $ct = x$ (and $ct' = x'$) in Figure 2.5.

Figure 2.6 shows the world lines for two particles travelling between the same two locations in spacetime. One particle has travelled at a constant speed (as represented by path A) while the other has taken a less direct route with varying speeds (path B). You will see in later sections of this chapter that, while world lines for objects can take any route through spacetime that does not involve travelling faster than the speed of light, the shortest route between two locations plays an important role in defining the geometry of spacetime.

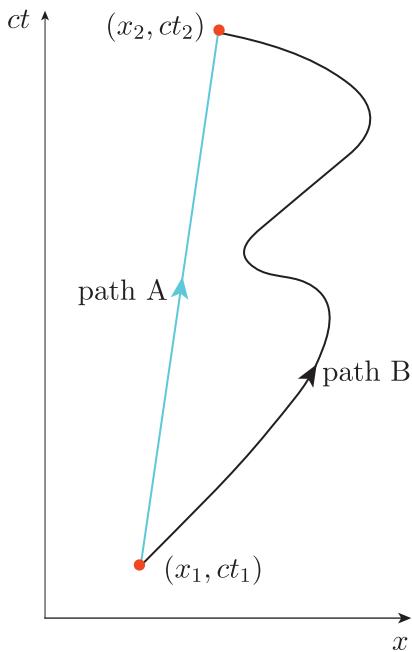


Figure 2.6 The world lines for two particles travelling between the same two points.

2.2.2 Causality and simultaneity

By visualising the relationship between events and how they are seen from different reference frames, spacetime diagrams make clear the significance of the finite speed of light. That the S' -axes converge around the line of $ct = x$ as relative speeds increase reflects the fact that the behaviour of light does not alter between the two reference frames. This has implications for causality and the connection between different events.

Figure 2.7 illustrates the concept of a **light cone**, via a spacetime diagram to which an extra spatial axis (y) has been added to make a three-dimensional representation. If we take the coordinate origin to be the location of a particular event, then the time axis extends both into the past (lower half of the plot, below the x -axis) and future (upper half) relative to that event. The shaded cone area encompasses all events that could be **causally connected** to the event at the origin.

- Consider two events, A (x_A, ct_A) , and B (x_B, ct_B) ; for example, A could be an explosion, and B could be a window shattering. What condition needs to be met for A to have caused B?
- Event A could only be the cause of B if some type of ‘information’, such as a transfer of force, can travel between the two locations x_A and x_B within the time interval $\Delta t = t_B - t_A$. The required speed of information transfer, $\Delta x/\Delta t$, must be less than c , because information cannot travel faster than the speed of light.

The edges of a light cone have a slope of $c\Delta t/\Delta x = 1$, because world lines with slopes < 1 would correspond to motion at speeds faster than c . This

means that there is no path by which information could travel between the origin of the spacetime diagram and a location outside the light cone unless it was travelling faster than c . Hence, if an event is within the light cone of event A, it may be causally connected to A; if not, then the two events cannot be connected.

- Is it possible for observers in different reference frames to disagree about whether events are causally connected?
- No. It would be impossible to develop logically consistent laws of physics if one event could cause another according to some observers, but not to others. This logical contradiction is avoided because all observers measure the speed of light as having the same value: the $ct = x$ line is the same in all frames, and therefore the light cone is too.

One reason that it is important to consider the question of causality in special relativity is because observers *can* disagree about the *order* of events in some situations, as the following example demonstrates.

Example 2.4

Figure 2.8 shows five events, A–E, as observed from two reference frames.

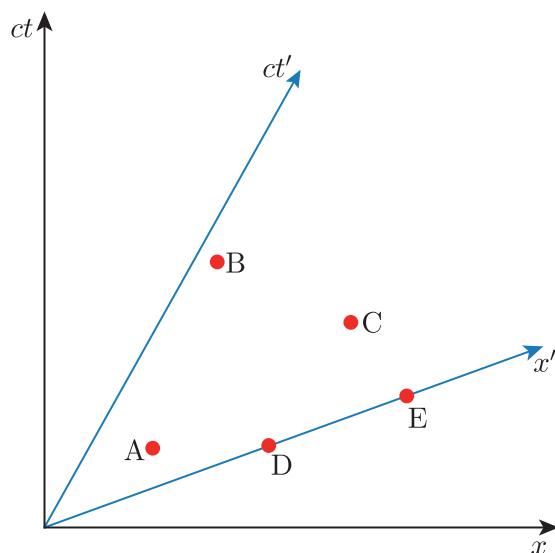


Figure 2.8 Five events plotted on a spacetime diagram representing two reference frames.

State the order in which the events occur according to: (a) an observer in frame S; (b) an observer in frame S'.

Solution

- In frame S, the events are observed to occur in a sequence defined by increasing values of ct : A = D (simultaneous), E, C, B.
- In frame S', where the time coordinate can be read off in the direction perpendicular to the x' -axis, the order is: D = E (simultaneous), A, C, B.

In other words, the observers will disagree about which events are simultaneous, and the ordering of some events. This situation is referred to as the **relativity of simultaneity**.

How can Example 2.4 be reconciled with the idea that observers must agree about whether events are causally connected? In the example above, it may at first glance look as though an observer in frame S could think that event A causes event E, whereas an observer in S' would think this cause and effect to be the other way round.

In fact, there is a straightforward solution. In Figure 2.8, a line connecting events A and E would have a very shallow slope, so any information travelling from A to E would need to cover a large x interval in a very small time interval – it would need to travel faster than the speed of light, so events A and E do not fall inside each other's respective light cones and so are not causally connected.

More generally, it turns out that situations in which observers disagree about the order of two events only *ever* arise where the two events are not contained within each other's light cone, so all observers will agree that one event cannot be the cause of the other (i.e. they will agree that it would require faster-than-light travel for information from one event to arrive at the location of the other before it happens). Therefore, provided information cannot travel faster than the speed of light, special relativity does not require us to abandon our sense of the reality of cause and effect.

2.2.3 Metrics in space and spacetime

In the chapter so far we have referred to ‘distances’ between events in a somewhat imprecise way. To extend our understanding of spacetime to encompass the behaviour of gravity and the expansion of the Universe, we need to define formally the spacetime separation between events. Our starting point is the concept of a **line element**: a small separation between coordinate locations. It is simplest to first consider the geometry of line elements in two and three dimensions, before coming back to four-dimensional spacetime.

Differential notation

In the previous sections we used notation such as Δx and Δt to indicate intervals in space and time. The mathematics of line elements (a form of differential geometry) is based on considering *infinitesimally* small coordinate intervals, and so from this point onwards the module materials will use the differential notation of ‘d’ to indicate such intervals (so dx and dt in the example of space and time).

The next example provides some practice in working with line elements.

Example 2.5

In two-dimensional space, the line element length dl shown in Figure 2.9 can be expressed in terms of the infinitesimal separations dx and dy , such that

$$dl^2 = dx^2 + dy^2 \quad (2.9)$$

By writing expressions for a coordinate location (x,y) in terms of the plane polar coordinates r and θ and then differentiating them, find an expression for dl^2 in terms of dr and $d\theta$.

Solution

First, write x and y in polar coordinates as

$$x = r \cos \theta$$

$$y = r \sin \theta$$

Using the total derivative rule, we can find expressions for dx and dy :

$$dx = dr \cos \theta - r \sin \theta d\theta$$

$$dy = dr \sin \theta + r \cos \theta d\theta$$

Squaring each expression and substituting into the original equation for dl^2 (Equation 2.9) gives:

$$\begin{aligned} dl^2 &= dr^2 \cos^2 \theta - 2r \cos \theta \sin \theta dr d\theta + r^2 \sin^2 \theta d\theta^2 \\ &\quad + dr^2 \sin^2 \theta + 2r \cos \theta \sin \theta dr d\theta + r^2 \cos^2 \theta d\theta^2 \end{aligned}$$

Recalling that $\sin^2 \theta + \cos^2 \theta = 1$, and noting that the two terms with $dr d\theta$ in them cancel out, this simplifies to:

$$dl^2 = dr^2 + r^2 d\theta^2 \quad (2.10)$$

The idea of a line element can be extended to three and four dimensions, and is related to a definition of the **spacetime separation** between two events, Δs , as follows:

$$\Delta s^2 = (c\Delta t)^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 \quad (2.11)$$

This definition looks similar to the equation for a simple three-dimensional (3D) distance, but with an additional term relating to how time intervals change, and with the spatial distances having negative sign. This negative sign ensures that Δs is invariant under Lorentz transformations.[†] All observers will agree about the spacetime separation between two events, which makes it a useful universal description of the geometry of spacetime.

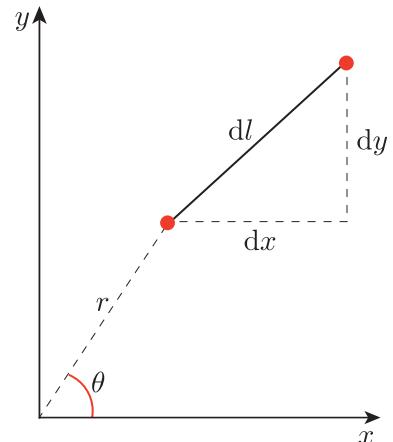


Figure 2.9 A line element dl in two-dimensional space, with x - and y -coordinate axes. The polar coordinates (r, θ) , corresponding to one end of the line element, are also shown.

The total derivative rule says that for any function $f(r, \theta)$

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \theta} d\theta,$$

where r and θ could be any independent variables.

[†]Mathematically, it doesn't matter whether the negative sign is applied to the time component or the spatial one; you may find examples outside the module material in which the signs of all of the terms in Equation 2.11 are reversed – this can be a source of confusion, so it is important to check which convention is being used.

When infinitesimal intervals in spacetime are considered, the spacetime separation becomes:

$$ds^2 = (c dt)^2 - dx^2 - dy^2 - dz^2 \quad (2.12)$$

Exercise 2.3

In the case where $dy = 0$ and $dz = 0$, show that the spacetime separation ds is invariant under Lorentz transformations (i.e. show that $ds = ds'$).

The expression for spacetime separation in Equation 2.12 is an example of a **metric** – a mathematical formulation of the relationship between coordinate intervals in a particular geometry. Equation 2.12 describes the **Minkowski metric**, which is the first of several metrics you will meet in this module. Metrics can also be written in the form of a summation:

$$ds^2 = \sum_{\mu, \nu=i}^n g_{\mu\nu} dx^\mu dx^\nu \quad (2.13)$$

where for 4D spacetime $i = 0$ and $n = 3$, and dx^0, dx^1, dx^2, dx^3 are the four-dimensional components of a vector $dx^\mu = (cdt, dx, dy, dz)$. This expression makes use of a tensor: a mathematical object that provides a concise way of manipulating multidimensional geometric relationships. The term $g_{\mu\nu}$ is known as the **metric tensor**,[‡] encapsulating the **metric coefficients** that apply to each coordinate. The metric tensor can be represented as a matrix with dimensions μ and ν . In the case of the Minkowski metric (Equation 2.12) this is written as:

$$g_{\mu\nu} = \begin{pmatrix} g_{00} & g_{01} & g_{02} & g_{03} \\ g_{10} & g_{11} & g_{12} & g_{13} \\ g_{20} & g_{21} & g_{22} & g_{23} \\ g_{30} & g_{31} & g_{32} & g_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad (2.14)$$

The non-zero diagonal elements of the metric tensor (i.e. g_{00}, g_{11}, \dots) are the coefficients of the coordinates for the ds^2 interval – compare them with the factors by which the coordinates $(cdt)^2, dx^2, dy^2$ and dz^2 are multiplied in Equation 2.12. The non-diagonal elements are all zero, because the metric doesn't include any terms involving (for example) $dx dy$.

The Minkowski metric, set out in two different forms by Equations 2.12 and 2.14, is a description of the geometry of space and time (in locally inertial frames). It enables calculations and measurements to be made about the properties and relations of lines, surfaces and volumes. Many other possible metrics exist. Note that it is conventional to use the index '0' for the time coordinate only, and so the summation in Equation 2.13 conventionally runs from 1 to n for an n -dimensional spatial metric.

[‡]Unfortunately, the tensor notation used in Equation 2.13 is easily confused with exponents when applied to individual elements – e.g. here dx^2 does not mean the square of dx , but the x -coordinate element with label '2'. A full introduction to tensors is beyond the scope of this module.

2.3 Curved space and spacetime

The ability to map the behaviour of space, or spacetime, using metric descriptions like those you were introduced to in the previous section is essential for astrophysics and cosmology. A metric helps to define the shortest distance between any two locations, which is important for understanding how light travels across the Universe.

Although the Minkowski metric applies to many useful situations, it is not the only metric we need to consider if we want to be able to tackle the subject of cosmology. The key insight of Einstein's theory of general relativity was a connection between gravity and geometry. To describe the Universe and its expansion (including the behaviour of light under the influence of gravity), it is necessary to extend our descriptions of geometry and metrics to encompass **curvature** of space and spacetime.

2.3.1 Flat and curved geometries

To introduce curved (i.e. **non-Euclidean**) geometries we start by taking a step back from the discussion of spacetime and consider further the geometry of two- and three-dimensional spaces. In the mathematical field of differential geometry, the type of smoothly varying spaces we will explore are referred to as **manifolds** – we will use both ‘manifold’ and ‘geometric space’ interchangeably.

In three dimensions, we have an intuitive idea of the difference between a flat and a curved surface. A piece of paper on a desk has a flat geometry. The surface of a balloon or a globe cannot be flattened in such a way as to lie flat on a desk and still retain the original distances between points and angles between lines.

- Does the surface of a cylinder have a flat or curved geometry?
- The surface of a cylinder has a flat geometry. A cylindrical tube can be unrolled to form a flat surface without distorting the relationships between points on its surface. An intrinsically curved surface such as the surface of a sphere cannot be flattened in this way (e.g. consider how maps of the world become distorted in flat projections).

Whether a particular surface (manifold) is flat or curved can be determined mathematically. To explore this, we will consider the metric that describes the surface of a sphere. Spherical coordinates are commonly used in physics because they simplify the mathematics of problems that are spherically symmetric.

Figure 2.10 shows how the three-dimensional spherical coordinates r , θ and ϕ are defined. They are related to x , y and z (Cartesian coordinates) as follows: $x = r \sin \theta \cos \phi$; $y = r \sin \theta \sin \phi$; and $z = r \cos \theta$.

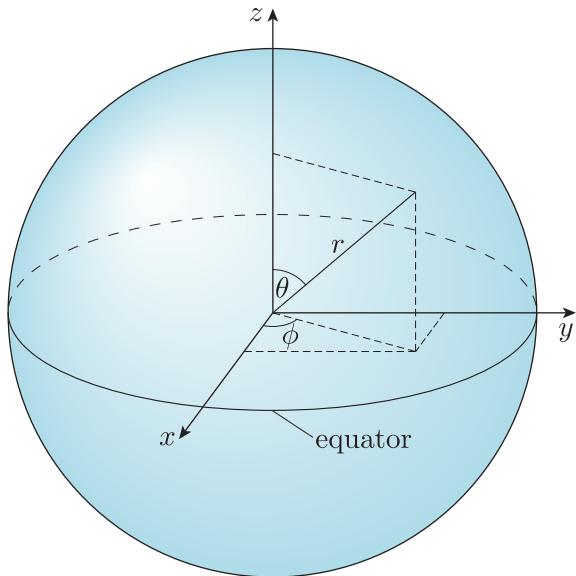


Figure 2.10 The 3D relation between spherical and Cartesian coordinates.

The length of a line element in three-dimensional space is given by $dl^2 = dx^2 + dy^2 + dz^2$. Using the relations above it is possible to show that this is equivalent to:

$$dl^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (2.15)$$

The surface of a sphere corresponds to the situation $r = R$, where R is a constant. In other words, we consider locations on a surface to be at a fixed distance R from the origin, as shown in Figure 2.11.

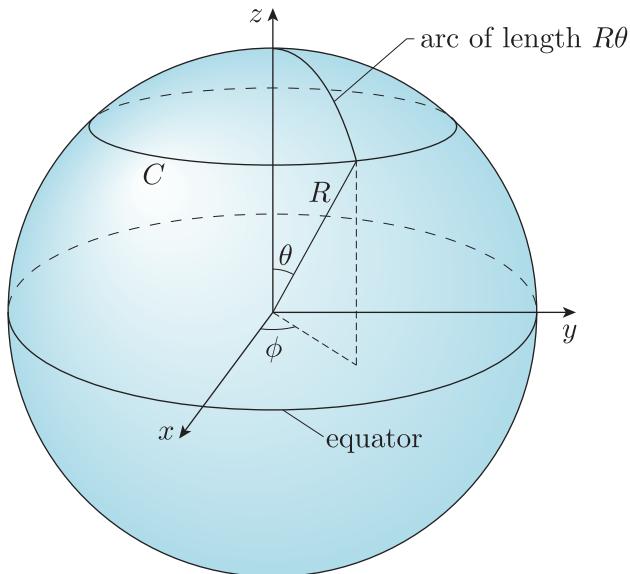


Figure 2.11 A circle C of constant θ coordinate drawn on the surface of a sphere.

In this geometry, $dr = 0$ (because $r = R = \text{constant}$), and so the metric for this geometry can be described by the line element:

$$dl^2 = R^2 d\theta^2 + R^2 \sin^2 \theta d\phi^2 \quad (2.16)$$

- How many dimensions does the space described by the metric in Equation 2.16 have?
- It describes a 2D space – the two variables θ and ϕ are the only dimensions. The third spatial dimension (the freedom to move in the r direction) does not exist in this geometry.

In the following example you will consider how the geometry described by Equation 2.16 affects the properties of circles.

Example 2.6

Consider the circle, C , drawn on the upper half of the surface of the sphere in Figure 2.11, which has its centre at the ‘north pole’. Every point on the circle has the same value of θ . Use the concept of a line element and the metric of the spherical surface to derive an expression for the circumference of the circle, C .

Solution

Because θ is constant, $d\theta = 0$, and so the line element that represents an infinitesimal distance on the sphere’s surface reduces to:

$$dl^2 = R^2 \sin^2 \theta d\phi^2$$

The circumference of C is the sum of all of the infinitesimal distances around the circle, so is given by the integral of the line element summed over all values of ϕ (the angle around the circle), which (in radians) ranges from 0 to 2π . This gives:

$$C = \int_0^{2\pi} dl = \int_0^{2\pi} R \sin \theta d\phi = R \sin \theta [\phi]_0^{2\pi} = 2\pi R \sin \theta$$

- Considering the same circle C , defined on the surface of the sphere shown in Figure 2.11, what is its radius?
- Remembering that the circle is defined *on* the curved surface, with its centre at the ‘north pole’, the radius is the length of the arc shown in the figure, i.e. $R\theta$.

If the surface had the same flat geometry as a plane, we would expect the circumference to be $2\pi \times \text{radius}$, i.e. $2\pi R\theta$, whereas the example found a circumference of $2\pi R \sin \theta$. So the circumference on the curved surface is *smaller* than that of a circle in flat space. This makes intuitive sense if we consider Figure 2.11 again, focusing now on the circle around the equator. Its circumference is simply $2\pi R$ (because $\theta = \pi/2$, and so $\sin \theta = 1$). But the circle’s radius, as drawn on the surface, is the arc from the ‘north pole’ to the equator, which is clearly much longer than R .

Hopefully, you are now persuaded that geometrical relationships differ in curved geometries. The reason for this is that the geometry determines the shortest distance between two points (known as a **geodesic**). In flat space the geodesic is always a straight line, but when movement is confined to a curved surface, the shortest route between two points on that surface always has to account for its curvature.

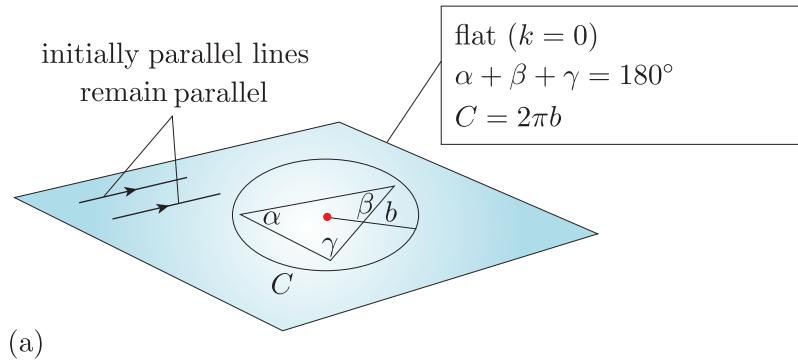
In the case of the surface of a sphere, shortest paths are always part of a **great circle**: a circle defined by the intersection of the surface with a plane that passes through the sphere's centre; for example, the Earth's Equator. Another example you may have come across is great-circle aeroplane flight paths, which trace the shortest route between two points on the Earth's surface, but can seem surprising when shown on a map that is a flat projection of the globe. (If you have studied observational astronomy previously, you will have encountered similar ideas in the context of astronomical coordinate systems defined on the celestial sphere.) The next exercise explores an example of distances in curved geometries.

Exercise 2.4

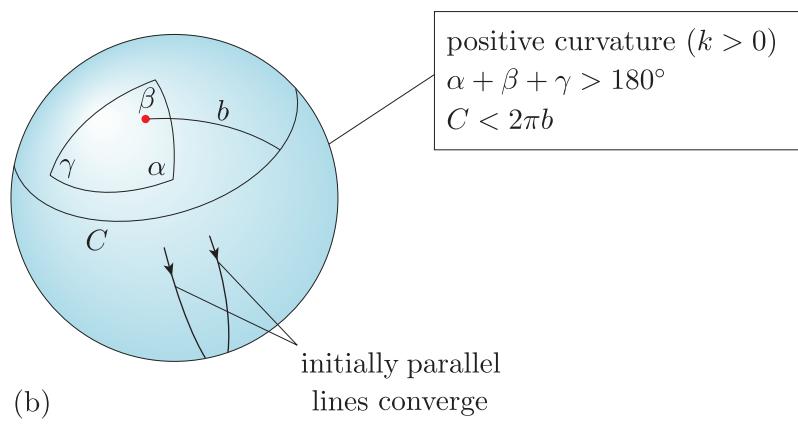
The cities of Calgary and Brussels are both at a latitude of 51° north, and are separated in longitude by 118° . The shortest (great-circle) distance between the two cities is ~ 7300 km, which involves a route that passes over Greenland, Iceland and the very north of Scotland. Calculate the distance between the two cities if, instead, a route that follows the 51° latitude line is taken (which would pass through eastern Canada and Southern England), and comment on how the two distances compare.

Figure 2.12 illustrates three ways in which curved 2D spaces – such as those of a spherical or a saddle-shaped surface – behave differently to flat ones. These differences concern: the relation between the circumference and the radius of circles (labelled C and b respectively in the diagram); the geometry of triangles; the behaviour of initially parallel lines. The figure also shows that it is possible to identify two different directions in which curved geometry can deviate from that of flat space – non-planar surfaces can have **positive curvature** or **negative curvature**, typically parameterised by a **curvature parameter**, k :

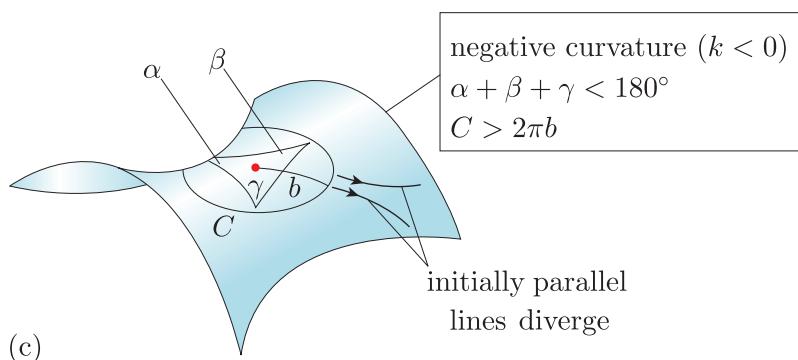
- In positively curved 2D spaces ($k > 0$), the interior angles of triangles sum to *greater than* 180° , circumferences of circles are smaller, and lines of shortest distance (geodesics) that are initially parallel may *converge*.
- In negatively curved 2D spaces ($k < 0$), the interior angles of triangles sum to *less than* 180° , circumferences of circles are larger, and initially parallel geodesics may *diverge*.



(a)



(b)



(c)

Figure 2.12 Differences in geometry of 2D spaces: (a) a flat space; (b) a space with positive curvature; (c) a space with negative curvature.

Similar deviations from Euclidean geometry occur for 3D and 4D curved spaces, and have important implications for the behaviour of light, and so for our understanding of spacetime in cosmology.

2.3.2 Defining and measuring curvature

How can curvature be measured? First consider a curved line, as shown in Figure 2.13.

- Use the geometric information in Figure 2.13 to determine whether the line is most curved at point A, B or C. Briefly explain your reasoning.
- The line is more curved at A than at B or C. At point A, the line deviates the most over a short distance from a straight tangent line, whereas at C the curve remains closest over a large distance to a straight tangent line.

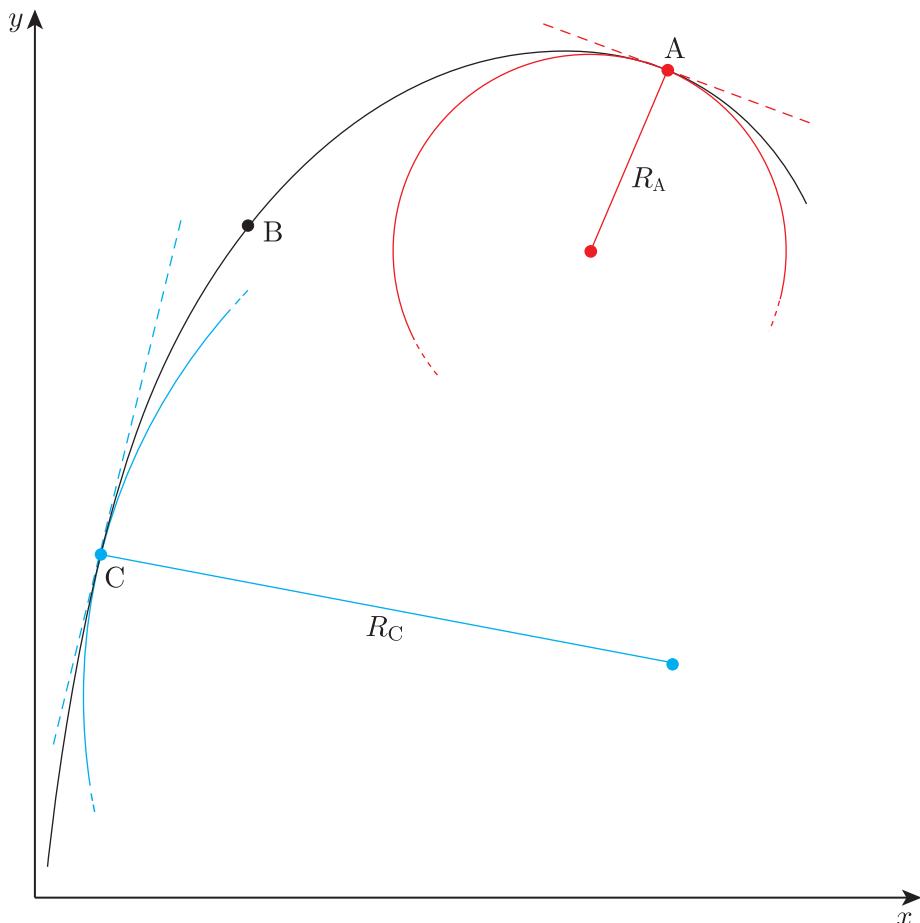


Figure 2.13 A line with varying curvature, marked with points A, B and C.

Figure 2.13 also shows that a line segment that is *more* curved can be approximated at a particular location by a segment of a circle of *smaller* radius, whereas a less curved line requires a circle of much larger radius to approximate its shape (compare the lengths of lines R_A and R_C). This leads to the curvature k_x , at a point x , being defined as the inverse of the radius of a circle that best matches the curve at that location:

$$k_x = \frac{1}{R_x} \quad (2.17)$$

Exercise 2.5

- (a) What shape is described by a line of constant curvature $k_x = 0.2 \text{ cm}^{-1}$?
- (b) What is the curvature k_x of a straight line (measured at any point, x , along it)?

For cosmology, the curvatures of both three-dimensional spaces and four-dimensional spacetime are relevant. These properties influence the paths taken by light and the distances between objects such as galaxies as the Universe expands. We therefore need to define curvature in a way that can be applied to a variety of geometries in multiple dimensions.

Figure 2.13 demonstrates that curvature is connected in some way to the behaviour of tangent lines. Where the curvature is high, the slopes of the tangent lines change rapidly as small steps are taken along the curve, whereas for low curvature the slopes do not change much as you move along the curve. In a 2D geometry, the slope is a first derivative of the coordinates, e.g. dy/dx , and so the rate at which the slope changes is a second derivative, e.g. d^2y/dx^2 .

Derivatives are quantities that connect how different coordinates change in relation to each other, and they can be used to construct a definition of curvature that is *intrinsic* to the geometry being measured. This means that the curvature can be determined without making measurements in a higher-dimensional space (in contrast to what is done in Figure 2.13, where the 1D curve is investigated with 2D (x,y) coordinates).

- What is the mathematical entity that describes the intrinsic geometry of a particular space (manifold)?
- The intrinsic geometry of a manifold is described by its metric.

Therefore, metric coefficients and their derivatives provide us with a way to determine intrinsic curvature. In a two-dimensional space, we can write the following (non-intuitive!) expression for intrinsic curvature, \mathcal{K} , in terms of metric coefficients and their derivatives:[§]

$$\begin{aligned}\mathcal{K} = & \frac{1}{4g_{11}g_{22}} \left[\frac{1}{g_{11}} \frac{dg_{11}}{dx^1} \frac{dg_{22}}{dx^1} + \frac{1}{g_{22}} \frac{dg_{11}}{dx^2} \frac{dg_{22}}{dx^2} + \frac{1}{g_{11}} \left(\frac{dg_{11}}{dx^2} \right)^2 + \frac{1}{g_{22}} \left(\frac{dg_{22}}{dx^1} \right)^2 \right] \\ & - \frac{1}{2g_{11}g_{22}} \left[\frac{d^2g_{22}}{(dx^1)^2} + \frac{d^2g_{11}}{(dx^2)^2} \right]\end{aligned}\quad (2.18)$$

where $g_{\mu\nu}$ refers to the metric elements as defined in Equation 2.13 and its ‘1’ and ‘2’ labels refer to coordinates (e.g. $x^1 = x$ and $x^2 = y$ for a 2D Cartesian geometry).

[§]Note again the potential source of confusion that in this context ‘ dx^1 ’ and ‘ dx^2 ’ indicate derivatives with respect to co-ordinates x^1 and x^2 , but the superscript ‘2’s in the final two terms indicate second-order derivatives.

For a subset of 2D metrics in which g_{11} is constant, and g_{22} depends only on x^1 , Equation 2.18 reduces to a simpler form:

$$\mathcal{K} = \frac{1}{4g_{11}(g_{22})^2} \left(\frac{dg_{22}}{dx^1} \right)^2 - \frac{1}{2g_{11}g_{22}} \left[\frac{d^2g_{22}}{(dx^1)^2} \right] \quad (2.19)$$

- What are g_{11} and g_{22} for a 2D plane polar geometry?
 - Referring back to Example 2.5, $g_{11} = 1$ and is the metric coefficient for r (it is the multiplier of the dr^2 term in the line element form of the metric, Equation 2.10), and $g_{22} = r^2$ and is the metric coefficient for θ (it is the multiplier of the $d\theta^2$ term in the same form of the metric).
-

Example 2.7

Find an expression for $\mathcal{K}(\theta, \phi)$ for the surface of a sphere, and show that the curvature does not depend on the coordinate location (θ, ϕ) – i.e. that the curvature of a spherical surface is the same at every location.

Solution

The metric for the surface of a sphere is given by Equation 2.16:

$$dl^2 = R^2 d\theta^2 + R^2 \sin^2 \theta d\phi^2$$

To calculate \mathcal{K} (using Equation 2.19) we need the two metric coefficients, g_{11} and g_{22} (where the ‘1’ and ‘2’ labels refer to θ and ϕ , respectively). From Equation 2.16, $g_{11} = R^2$ and $g_{22} = R^2 \sin^2 \theta$.

To evaluate the equation, we need to calculate the first and second derivatives of g_{22} with respect to x^1 , i.e. θ :

$$\frac{dg_{22}}{dx^1} = \frac{dg_{22}}{d\theta} = 2R^2 \sin \theta \cos \theta$$

and so

$$\frac{d^2g_{22}}{(dx^1)^2} = 2R^2(-\sin^2 \theta + \cos^2 \theta) = 2R^2(\cos^2 \theta - \sin^2 \theta)$$

Because Equation 2.19 is algebraically messy, we will first work out each of its two terms individually. Substituting the first derivative of g_{22} into the first term in the equation gives:

$$\frac{1}{4g_{11}(g_{22})^2} \left(\frac{dg_{22}}{dx^1} \right)^2 = \frac{1}{4R^2(R^2 \sin^2 \theta)^2} (2R^2 \sin \theta \cos \theta)^2$$

Doing the same for the second term, dependent on the second derivative of g_{22} , gives:

$$\frac{1}{2g_{11}g_{22}} \left[\frac{d^2g_{22}}{(dx^1)^2} \right] = \frac{1}{2R^2R^2 \sin^2 \theta} [2R^2(\cos^2 \theta - \sin^2 \theta)]$$

We can now use the two expressions above to find an expression for \mathcal{K} :

$$\begin{aligned}\mathcal{K}(\theta, \phi) &= \frac{1}{4R^2(R^2 \sin^2 \theta)^2} (2R^2 \sin \theta \cos \theta)^2 \\ &\quad - \frac{1}{2R^2 R^2 \sin^2 \theta} [2R^2 (\cos^2 \theta - \sin^2 \theta)]\end{aligned}$$

which simplifies to

$$\begin{aligned}\mathcal{K}(\theta, \phi) &= \frac{\cos^2 \theta}{R^2 \sin^2 \theta} - \left(\frac{\cos^2 \theta - \sin^2 \theta}{R^2 \sin^2 \theta} \right) \\ &= \frac{\cos^2 \theta}{R^2 \sin^2 \theta} - \frac{\cos^2 \theta}{R^2 \sin^2 \theta} + \frac{\sin^2 \theta}{R^2 \sin^2 \theta}\end{aligned}$$

The first two terms cancel out, and the final one simplifies further to give

$$\mathcal{K} = \frac{1}{R^2}$$

You can see that – in the end, after much algebra – the curvature of the spherical surface does not depend on the location (θ, ϕ) position). For the surface of a sphere of particular radius R , the curvature is the same everywhere, and depends only on the sphere's radius.

Example 2.7 was intended to give you a flavour of the mathematical and physical meaning of spatial curvature. Curvature becomes even more complex to calculate for higher-dimensional spaces. \mathcal{K} is a simplified form of the **Riemann curvature tensor**, $R_{\mu\rho\nu}^\lambda$, sometimes just referred to as the Riemann tensor. The Riemann tensor is a set of equations analogous to Equation 2.18 that encapsulates the curvature at any location in a multidimensional space.

Online resources: Riemannian geometry

The mathematical toolkit that underpins the theory of general relativity is Riemannian geometry. Our primary aim in this module is to develop and apply cosmological models, and so – although it is a fascinating topic – introducing the wealth of definitions and terminology required for a full discussion of Riemannian geometry is beyond our scope. The online resources for this chapter provide some additional information about the subject for interest only.

The crucial point to take away, and the reason we have shown you the somewhat complicated form for the two-dimensional case in Equation 2.18, is that, like \mathcal{K} , the curvature tensor is constructed entirely from *derivatives of the metric coefficients*. You will not be asked to calculate or manipulate it directly.

The essential concepts to remember from this section are summarised in the following box.

Curvature of geometric spaces

- Curvature affects the shortest distance between two points in space.
- Geometrical relationships, such as the properties of circles and triangles, differ in flat and curved spaces.
- Curvature is an *intrinsic* property of a geometric space, which can be determined at a given location in space if the metric is known.
- Intrinsic curvature is determined from a combination of derivatives of the metric coefficients, given for 2D geometries by Equation 2.18.
- The Riemann curvature tensor is a more complex combination of derivatives of metric coefficients. It encapsulates the curvature at any location of a manifold of arbitrary dimensions (and so can be used to quantify curvature in the 3D and 4D situations relevant for general relativity and cosmology).

So far, we have considered only spatial dimensions, but you have seen that special relativity requires us to work with four-dimensional spacetime, in which spatial and time dimensions can become intertwined. In the next section we consider how the concepts of curvature explored here can be applied to 4D spacetime.

2.3.3 Curved spacetime and geodesics

Curvature is an important concept in four-dimensional spacetime as well as in the spatial geometries discussed so far. Although, formally, spacetime metrics are termed ‘pseudo-Riemannian’ (a distinction not important for this module), the Riemann tensor and the arguments about curved geometries in the previous section can be applied to spacetime metrics as well as to purely spatial geometries.

- Curvature depends on combinations of the derivatives of metric components with respect to different coordinates. Considering the Minkowski metric (Equation 2.12), what basic statement must be true about any derivatives of its components?
- All of the components of the Minkowski metric are constants (either 1 or -1), and so their derivatives with respect to another coordinate must be zero.

Because all of the partial derivatives of the components of its metric must be zero, we can conclude that Minkowski spacetime has zero curvature: it is a flat geometry. However, it is possible to construct spacetime geometries that are curved, and you will see in the next chapter that such geometries are a consequence of the theory of general relativity.

Another concept from this chapter that is important to general relativity is the geodesic: the path of shortest distance between two points for a particular metric (where distance is the spacetime invariant distance ds).

Geodesics have the form of straight lines in flat (Minkowski) spacetime, but take different forms in curved geometries, such as the great circles around the Earth's surface. Geodesics can also be divided into three categories, depending on the nature of the spacetime separation (or metric interval) between the two points considered.

- **Null geodesics** are curves for which $ds^2 = 0$ for every interval along the curve. Only particles travelling at the speed of light (i.e. photons) can travel along null geodesics, which are also known as 'light-like' geodesics.
- **Time-like geodesics** are curves where $ds^2 > 0$ for every interval along the curve. The world lines of particles (moving at $v < c$) must follow time-like paths.
- **Space-like geodesics** are curves where $ds^2 < 0$ for every interval along the curve. These correspond to paths between events that are outside one another's light cones – particles (including photons) cannot travel along a space-like path.
- Can observers logically disagree about the order of two events that are connected by a geodesic that is: (i) null, (ii) time-like, and (iii) space-like?
- Observers must agree about the order of events in situations where one event could be the physical cause of another. Therefore, two events that can be connected by a null or a time-like geodesic must have a fixed order for all observers. Observers may disagree, however, about the order of events for which the separation is space-like because there can be no causal relationship between them if neither event is inside the other's light cone.

The world lines along which particles travel are always time-like. The earlier discussion of special relativity implies that – whether in flat or curved spacetime – observers in different reference frames may disagree about how quickly time elapses along those paths, i.e. how changes in the t -coordinate along a path relate to changes in the other coordinates. In the next section you will meet a definition of time that all observers in a region described by the same metric can agree on – a concept that will be essential for building geometric models of the Universe.

2.3.4 Proper time

Metrics are defined such that all observers can agree on the invariant spacetime interval between two events, ds . An invariant time interval, called the **proper time**, $d\tau$, can be defined in relation to the spacetime interval, ds , in a very simple way:

$$d\tau = ds/c \quad (2.20)$$

Because ds and c are both the same for all observers, $d\tau$ is a measure of time that everyone can agree on.

It turns out that proper time has another useful property, shown in the following example.

Example 2.8

Consider a particle travelling along a time-like geodesic in Minkowski (i.e. flat) space. An observer who is not travelling with the particle sees it move a very short distance between two locations that have coordinates of (x_A, y_A, z_A) and (x_B, y_B, z_B) in the observer's reference frame.

Write down expressions for the proper time interval that elapses when the particle travels between these two locations, both in the reference frame of the observer and in the reference frame of the particle. Comment on how proper time relates to how time is perceived in the reference frame of a moving object.

Solution

We start by writing down expressions for the interval ds in the two reference frames. Firstly, for that of the observer:

$$ds^2 = c^2 dt^2 - (x_B - x_A)^2 - (y_B - y_A)^2 - (z_B - z_A)^2$$

where we have used Equation 2.12 (the Minkowski metric), and substituted the individual coordinate separations between events into the spatial separation terms.

For the reference frame of the particle we will use primed notation to denote the different reference frame:

$$ds'^2 = c^2 dt'^2 - dx'^2 - dy'^2 - dz'^2$$

We can immediately simplify this expression hugely because, in the reference frame attached to the moving particle, its spatial coordinates do not change at all: it is at rest. So

$$ds'^2 = c^2 dt'^2$$

If we now apply Equation 2.20 to the two expressions, we find that

$$d\tau^2 = dt^2 - [(x_B - x_A)^2 - (y_B - y_A)^2 - (z_B - z_A)^2]/c^2$$

and

$$d\tau'^2 = dt'^2$$

The spacetime separation is the same for both frames (i.e. $ds = ds'$), and so from the definition of proper time we can conclude that $d\tau = d\tau' = dt'$.

In other words, for a time-like world line, the proper time interval between two events on that world line is equivalent to the coordinate time interval dt measured in the frame at rest in relation to those events.

The example provides a very useful result, which can be generalised to apply to curved as well as flat spacetimes.

Proper time and coordinate time

For events on the world line of a moving observer – meaning that they all occur at the same, unchanging spatial coordinate location in the observer’s frame – the proper time corresponds to the coordinate time that would be measured by a clock travelling with the observer.

For an observer in a different reference frame, the proper time between events does not correspond to the coordinate time, but can be determined from the metric based on the spatial coordinates of those events.

The concept of proper time, and its relation to coordinate time, will be important for building and interpreting metric descriptions of curved and flat spacetime in different regions of the Universe. This is the topic of the next chapter.

2.4 Summary of Chapter 2

- Measurements of intervals in space and time, and related quantities such as velocity, depend on the **reference frame** of the observer who is measuring them.
- The theory of special relativity enables a consistent framework for applying the laws of physics in all **inertial frames**, in which the **Lorentz transformations** can be used to relate measurements of distance and time **intervals** in different frames:

$$\begin{aligned}\Delta x' &= \gamma(\Delta x - V\Delta t) \\ \Delta y' &= \Delta y \\ \Delta z' &= \Delta z \\ \Delta t' &= \gamma(\Delta t - V\Delta x/c^2)\end{aligned}$$

where the **Lorentz factor**

$$\gamma = \frac{1}{\sqrt{1 - V^2/c^2}} \quad (\text{Eqn 2.5})$$

- Where two reference frames are moving at different speeds from one another, the time interval between two **events** observed to occur at the same location in one frame will be measured as different in the other frame, and this **time dilation** is an example of the consequences of special relativity.
- The relationship between events and the **world lines** representing the paths of moving objects can be investigated using **spacetime diagrams**.
- **Light cones** provide a way of conceptualising which events can be **causally connected** to each other. If two events A and B take place, then event A cannot have been caused by event B if B occurs at a location outside A’s light cone.

- **Metrics** define the geometric relationship between distance and time intervals, with the **Minkowski metric** defining a **spacetime separation** on which all observers can agree:

$$ds^2 = (c dt)^2 - dx^2 - dy^2 - dz^2 \quad (\text{Eqn 2.12})$$

- More generally, the **metric coefficients** of a geometric space with n dimensions can be defined by

$$ds^2 = \sum_{\mu, \nu=i}^n g_{\mu\nu} dx^\mu dx^\nu \quad (\text{Eqn 2.13})$$

where $g_{\mu\nu}$ is the **metric tensor** and x^μ are the coordinates for each dimension, which are conventionally labelled from 1 to n for an n -dimensional space, and from 0 to 3 for 4-dimensional spacetime.

- **Curvature** of space is an intrinsic property of a particular metric tensor (or just ‘metric’). It affects the paths of shortest distance between points, as well as geometric relationships such as the properties of circles and triangles.
- The curvature of a 2D surface where g_{11} is constant and g_{22} depends only on x^1 can be calculated from the metric coefficients according to:

$$\mathcal{K} = \frac{1}{4g_{11}(g_{22})^2} \left(\frac{dg_{22}}{dx^1} \right)^2 - \frac{1}{2g_{11}g_{22}} \left[\frac{d^2g_{22}}{(dx^1)^2} \right] \quad (\text{Eqn 2.19})$$

- The **Riemann curvature tensor** is a more complex combination of derivatives of the metric. It provides a rigorous definition of curvature that can be applied to any multidimensional **manifold**.
- A **geodesic** is the path of shortest distance between two points for a particular metric. Geodesics can be classed as **null** ($ds^2 = 0$), **time-like** ($ds^2 > 0$) or **space-like** ($ds^2 < 0$).
- **Proper time** is an invariant measure of time intervals between events. It is defined as

$$d\tau = ds/c \quad (\text{Eqn 2.20})$$

and corresponds to the coordinate time interval measured by an observer who is co-located with both events.

Chapter 3 The geometry of the Universe

The previous chapter demonstrated that any physical theory of the Universe must account for the interlinked nature of space and time, and introduced some geometric methods for describing spacetime. This chapter puts those ideas together with a crucial insight about the nature of gravity, which led to the development of the general theory of relativity.

In this chapter you will explore the ideas underpinning general relativity as well as some key evidence in support of the theory, before moving on to explore two of the most important metrics for understanding and observing the distant Universe: the Schwarzschild metric, which describes spacetime surrounding a massive object, and the Robertson–Walker metric, which describes the geometry of our expanding Universe.

Objectives

Working through this chapter will enable you to:

- explain the equivalence principle, and describe its significance for our understanding of gravity
- describe Einstein's field equations and understand and explain the meaning of the terms included in them
- summarise key evidence in support of the theory of general relativity
- describe the behaviour of spacetime in the vicinity of a massive object such as a planet or a black hole, and apply the Schwarzschild metric to solve problems in this situation
- understand and explain the form of the Robertson–Walker metric that describes an expanding Universe
- define and manipulate three key cosmological parameters linked to this metric: the scale factor, curvature parameter and Hubble parameter.

3.1 Gravity as geometry

3.1.1 Free fall and the equivalence principle

The extension of the theory of special relativity to encompass the behaviour of gravity began with Einstein's realisation that for an observer falling freely from a great height, accelerating downwards as a result of a gravitational field, physics behaves as though gravity is 'switched off'.

Figure 3.1 shows an observer in a lift falling down an airless lift shaft with no friction, who has let go of an object while falling freely. The downward accelerations of the lift, the observer and the object will all be the same, because a principle called the **universality of free fall** guarantees that acceleration under gravity is independent of an object's mass or

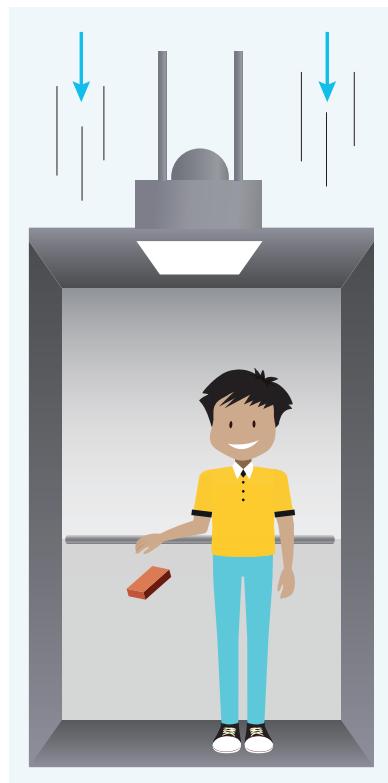


Figure 3.1 Effect of dropping an object in a freely falling lift.

composition. This means that the released object will appear to remain stationary to the observer; it will not fall to the floor of the lift, because the lift and the object are accelerating by the same amount. If the observer exerts a force on the object (e.g. pushing it sideways) they will see it behave according to Newton's laws of motion (e.g. being displaced horizontally in the direction of the applied force), so the freely falling observer's frame is acting like an inertial frame of reference.

- If the released object in the freely falling lift appears to 'float' rather than drop to the floor, should the same apply to the person?
- If the person was standing on the floor at the time the lift began freely falling, then they would remain in the same place relative to the top and bottom of the lift, as shown in Figure 3.1; similarly, if they raised their feet off the ground they would not drop back to the bottom of the lift.

In contrast to the situation shown in Figure 3.1, where the dropped object appeared *not* to fall from the perspective of the person in the lift, we can consider an observer located in a region where there is no gravitational field. Figure 3.2 shows what would happen if someone dropped an object while in a rocket that is undergoing uniform acceleration and is not subject to gravitational forces. If the uniform acceleration, a , has the same magnitude as g , the gravitational acceleration on the surface of the Earth, then the observer will see behaviour that appears identical to gravity on Earth – namely, the object will appear to fall. The relative motion between the dropped object and the floor of the rocket will be equivalent to that of a released object that falls to the ground on Earth, because their relative accelerations are the same in the two situations.

- How are these two situations equivalent if the rocket in Figure 3.2 is accelerating in an *upward* direction relative to the person inside it, while a person on the surface of the Earth feels a *downward* gravitational pull?
- A person of mass m on the surface of the Earth exerts a force on the Earth's surface of $F = -mg$, and so by Newton's third law the Earth's surface exerts an opposing force of the same magnitude, mg . In Figure 3.2, the floor of the accelerating rocket exerts an upward force of $F = ma$ on the person, so that if the magnitudes of a and g are the same, then the effect is equivalent.

This agreement between the physics in regions of uniform acceleration and in regions under gravity is known as the **equivalence principle**. This principle has a 'weak' and a 'strong' form. The weak equivalence principle refers specifically to the *motion* of objects and can be stated as follows.

The weak equivalence principle

Within a localised region of spacetime near to a concentration of mass, the motion of objects caused by gravitational effects alone cannot be distinguished by any experiment from the motion of objects within a region of appropriate uniform acceleration.

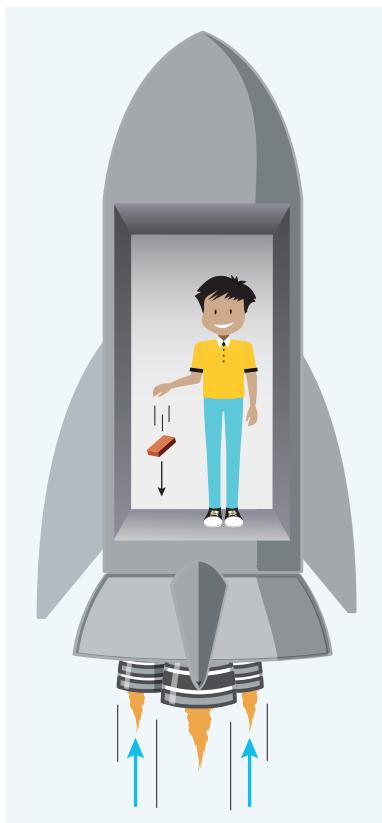


Figure 3.2 Effect of dropping an object in a rocket undergoing uniform acceleration.

It is important to note the reference to a *localised* region of space. Unlike the theory of special relativity, when thinking about general relativity we need to consider **local inertial frames**: the location of an observer matters, as well as their relative motion.

To see why this is the case, consider an observer in a lab on the surface of the Earth, who drops two objects separated by a certain horizontal distance. If the objects are falling towards the centre of the Earth under gravity, then the directions of their acceleration vectors will be very slightly different, as represented in Figure 3.3a. By contrast, for the equivalent situation we considered earlier – namely, an upwardly accelerating rocket that is *not* subject to a gravitational field – the directions in which the two objects fall are exactly the same, as shown in Figure 3.3b.

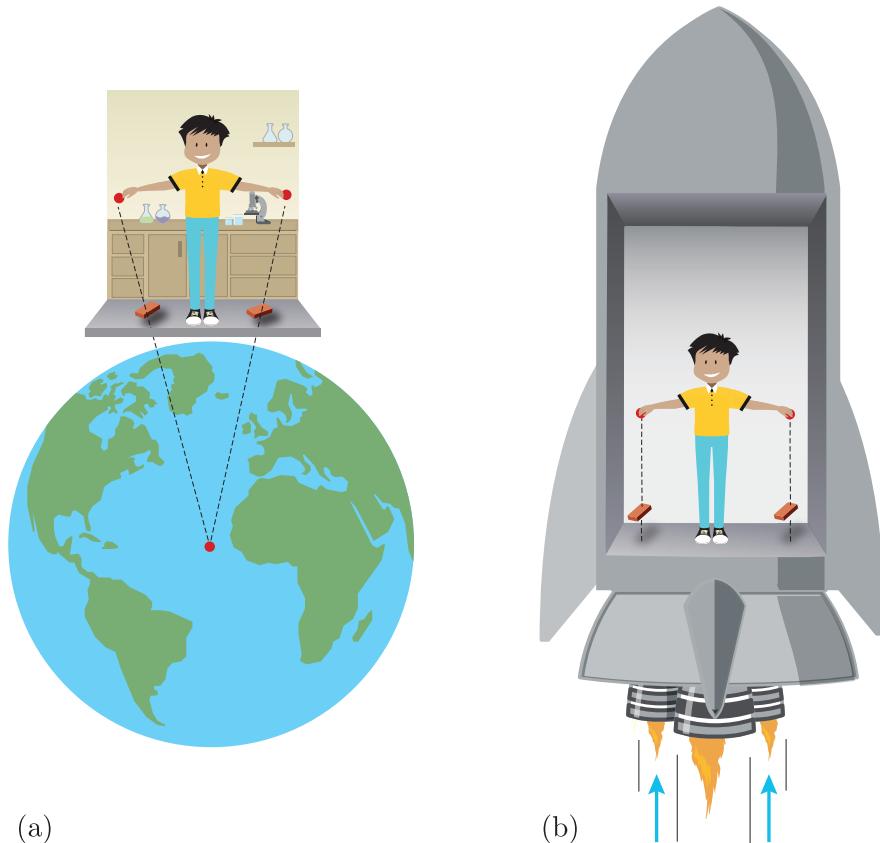


Figure 3.3 Dashed lines showing the motion of objects falling (a) under gravity on Earth and (b) in a rocket with uniform acceleration that is unaffected by gravitational forces. Differences in the motions of the objects become noticeable if the reference frames applied are insufficiently local.

Although the difference in the situation of Figure 3.3a would be undetectable in a lab-sized region, the slight deviation in the direction of acceleration from the exactly vertical would become measurable with an experiment if the reference frame, and the distance between the dropped objects, spanned a significant angular extent relative to the Earth’s centre. It is therefore *crucial* to always define reference frames that cover a region

and time interval that are sufficiently small as to be inertial within the accuracy needed to draw useful conclusions.

The weak equivalence principle has been very well tested experimentally, and is known to be accurate to an uncertainty level of less than 1 part in 10^{11} . The ‘strong’ form of the equivalence principle states that not just motion, but *any* physical behaviour of objects cannot be experimentally distinguished. This is less definitively proven, but any deviations must again be small. Much of the behaviour we will consider in the rest of the chapter is a consequence of the weak equivalence principle.

3.1.2 Einstein’s field equations

The equivalence principle led Einstein to the idea that a *metric theory of gravity* is possible. Under this model, whose postulates are summarised below, the motion of test particles that are subject to gravity is determined by the form of the metric in a particular location, instead of being caused by gravitational forces. More specifically, in general relativity, test particles (including light) will follow geodesics.

- What is a geodesic, and why might a test particle follow such a path?
- Geodesics are the shortest routes between two points, and so are likely to be the path requiring the least energy. Geodesic paths will differ in flat and curved geometries, as discussed in the previous chapter.

Postulates of a metric theory of gravity

- The geometric properties of spacetime are described by a metric that encapsulates the influence of gravity on spacetime.
- The world line of a test particle subject only to gravitational influence (i.e. not subject to electromagnetic or nuclear forces) is a geodesic of this metric.
- The world line of a light ray or other electromagnetic signal travelling in a vacuum is a null geodesic of this metric ($ds^2 = 0$).

Geodesics in Minkowski (flat) spacetime are straight lines, which is consistent with the way Newton’s first law describes the motion of bodies that are not subject to forces. The influence of gravity must change the form of the geodesics, and so cause spacetime to be curved. The general theory of relativity sets out Einstein’s theory of how the presence of matter determines the metric of four-dimensional spacetime.

More specifically, this theory is encapsulated in **Einstein’s field equations**, which describe in mathematical form the relationship between geometric curvature and the distribution of mass and energy. As noted in the previous chapter, it is beyond the scope of this module to teach the full mathematics of Riemannian geometry necessary to derive the field equations. Therefore, we can only introduce them and describe their

meaning in a qualitative way. This will be sufficient to allow us to use some important solutions to the field equations in order to study cosmology.

Einstein's field equations can be set out as in Equation 3.1, where G is the gravitational constant and the other terms are summarised in Table 3.1.

Einstein's field equations of general relativity

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\frac{8\pi G}{c^4}T_{\mu\nu} \quad (3.1)$$

The tensor quantities in Equation 3.1 have two indices, which cycle through each dimension of spacetime. This means that Equation 3.1 describes $2^4 = 16$ equations, which need to be solved jointly to find the metric for a particular energy–momentum distribution. In practice, only 10 of the equations are independent, so there are usually 10 equations to solve, not 16.

Table 3.1 The meaning of terms in Einstein's field equations

Quantity	Definition
$R_{\mu\nu}$	The Ricci curvature – this is a simplified form of the Riemann curvature tensor.
$g_{\mu\nu}$	The metric tensor.
R	The curvature scalar – a further simplification of the Ricci curvature.
$T_{\mu\nu}$	The energy–momentum tensor (or stress-energy tensor) – a set of terms describing the distribution and flow of energy and momentum caused by the presence of matter and radiation.

In other descriptions of general relativity you may sometimes see the left-hand side of Equation 3.1 written in simplified form as a single tensor $G_{\mu\nu}$, known as the **Einstein tensor**.

Exercise 3.1

In Chapter 1 you were introduced to a famous description of the theory of general relativity by John Wheeler: ‘Space tells matter how to move; matter tells space how to curve’ (Misner, Thorne and Wheeler, 1973, p. 5). Write a short explanation of which elements of the theory of general relativity as introduced in this chapter relate to each half of this description.

The field equations are a very concise way to encapsulate a complete description of how gravity behaves anywhere in the Universe, for any possible distribution of matter and energy. Solving the field equations is not an easy task, and you will not be asked to do it in the module. For situations other than the simplest distributions of mass it is usually necessary to use computational methods, rather than doing the algebra by hand.

Later in this chapter you will investigate two of the most important solutions of the field equations: the metric that describes spacetime near a single massive object (e.g. a planet or a star) and the metric that describes the expansion of the Universe as a whole. However, we will first briefly discuss how the theory of general relativity can be tested, and what evidence exists to convince us we can rely on it for studying cosmology.

3.1.3 Evidence for general relativity

A wide range of experiments have been conducted to test the theory of general relativity. To date, the theory has passed all of these tests successfully, with no evidence of discrepancies.

A famous early triumph for the theory was its ability to predict the **precession** of the perihelion of Mercury. Midway through the nineteenth century, it was already known that the orbit of the planet Mercury exhibited some unexplained behaviour: the planet's perihelion – the point of closest approach to the Sun – had been measured to shift position with each successive orbit, as illustrated in Figure 3.4a. This effect was not explained by Newtonian mechanics, but Einstein demonstrated that it is expected in general relativity, with the predicted angular shift being in very good agreement with what was observed in practice.

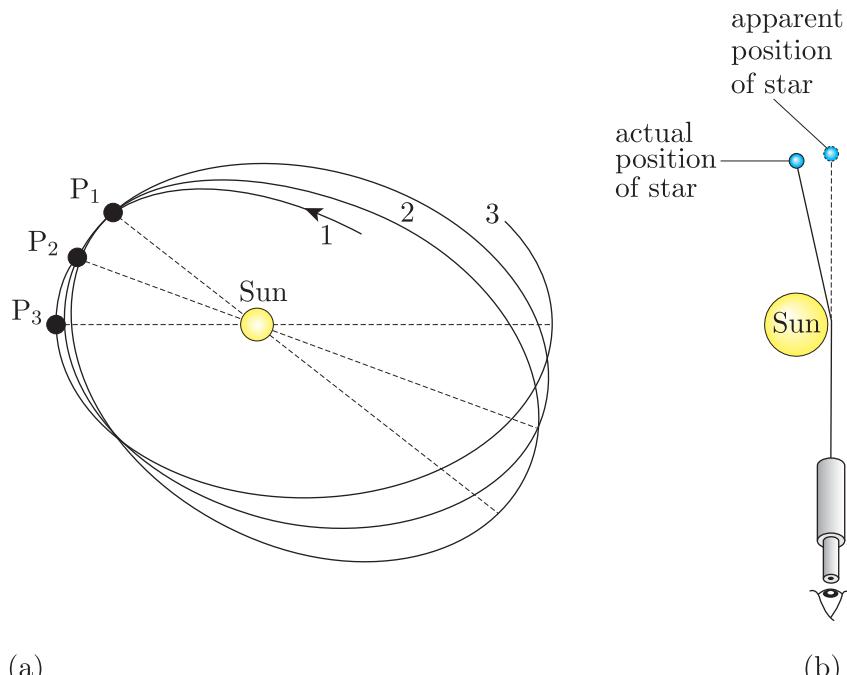


Figure 3.4 Two predictions of general relativity: (a) the changing orbital shape of Mercury with time, with the perihelion location P advancing with each successive orbit (labelled 1, 2, 3, ...); (b) the deflection of a star's light when the Sun passes near to the star's light path to Earth.

General relativity also predicts a comparatively large deflection in the path of light travelling from a distant star and passing near to the Sun, as

illustrated in Figure 3.4b. Testing this prediction requires making measurements of the positions of stars and recording how these change when the Sun passes very close to their projected location on the sky. Measuring the resulting deflections is tricky because of the brightness of the Sun, but is possible during a total eclipse.

Arthur Eddington carried out a series of such measurements in the early twentieth century, which helped to promote Einstein's theory and made international headlines (Figure 3.5). More recently, the predictions of light bending by the Sun have been verified to an accuracy of $< 0.04\%$ using observations of quasars via radio interferometry, a method of measuring radio waves to very high spatial precision.

Another famous experiment measured an effect known as gravitational redshift, which is a consequence of relativistic time dilation due to gravity, and is discussed further in Section 3.2.3. The **Pound–Rebka experiment** of 1960 confirmed that the Earth's gravity causes a detectable redshifting of light, and calculated the magnitude of this to be less than 1% with the general relativity prediction. The **Shapiro delay** is another important general relativistic effect, in which radar signals are delayed when they pass close to a massive object.

Modern astronomical observations also provide extensive evidence for general relativity. This includes the gravitational lensing of distant galaxies and quasars (the distortion of their appearance in images due to the path their light has taken to reach us) and the behaviour of neutron stars and black holes, which will be discussed further in the next section.

One of the most exciting recent discoveries in astrophysics was the direct detection in 2015 of **gravitational waves** with an instrument on Earth. General relativity predicts that moving sources of gravitation will emit waves that propagate as time-varying distortions in spacetime. Detections of such gravitational waves were made by the Laser Interferometer Gravitational-wave Observatory (LIGO) and the Virgo interferometer (located in the US and Italy, respectively), and led to the Nobel Prize in Physics being awarded to Thorne, Weiss and Barish in 2017. Gravitational waves had actually been detected more indirectly in the 1970s by radio monitoring of a binary pulsar system, in which two neutron stars are gradually spiralling closer together, with this shrinking orbit caused by the loss of energy via gravitational waves. This earlier discovery also received a Nobel Prize in Physics, which was awarded to Hulse and Taylor in 1993.

Figure 3.6 shows results from the two Nobel Prize-winning studies mentioned above. Panel (a) shows the first *direct* detection of gravitational waves. The data are plotted as a quantity called strain, which measures the relative distortion of spacetime in different directions caused by the passage of gravitational waves from distant events, such as mergers of black holes and neutron stars. The experimental data are presented alongside models that are used to infer the properties of the merging objects.

Panel (b) shows observations of the decay of the orbit of the binary pulsar system identified by Hulse and Taylor over 17 years. Red dots plot the

LIGHTS ALL ASKEW, IN THE HEAVENS

**Men of Science More or Less
Agog Over Results of Eclipse
Observations.**

EINSTEIN THEORY TRIUMPHS

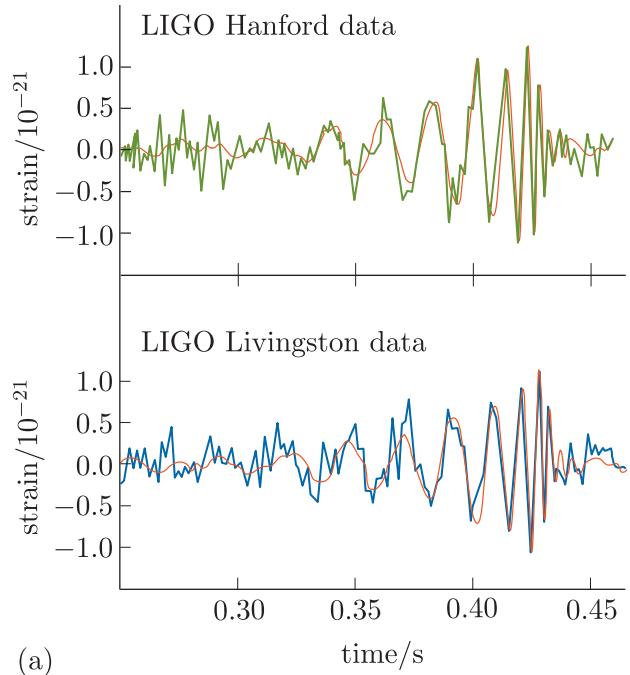
**Stars Not Where They Seemed
or Were Calculated to be,
but Nobody Need Worry.**

A BOOK FOR 12 WISE MEN

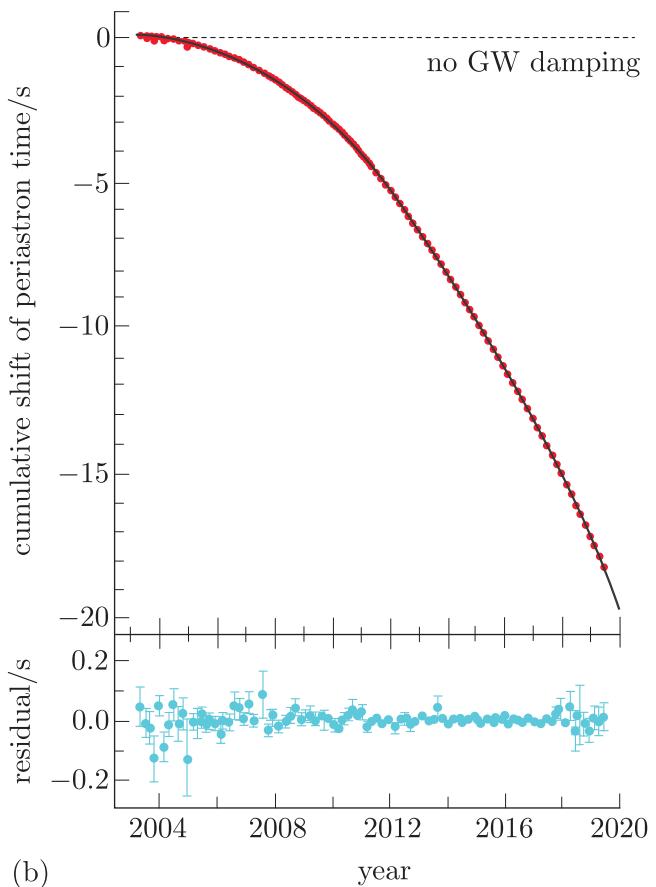
**No More in All the World Could
Comprehend It, Said Einstein When
His Daring Publishers Accepted It.**

Figure 3.5 A *New York Times* newspaper headline describing one of Eddington's solar eclipse expeditions to test general relativity using the light-deflection method.

shift in the time of periastron (closest approach) compared with the orbital change predicted by general relativity (solid black curve beneath the dots). The lower panel highlights the close alignment between the data and the predicted model.



(a)



(b)

Figure 3.6 (a) The first gravitational wave signals, measured independently at the LIGO Hanford and Livingston detectors, compared with a black-hole merger model (red-orange lines). (b) Radio monitoring of the orbital decay of the Hulse–Taylor binary pulsar system (red dots) plotted against the prediction of general relativity (black curve). The horizontal dashed line shows the expected behaviour if no gravitational waves (GW) were transporting energy away; the lower panel highlights the close alignment between the data and the predicted model; the lower plot shows the residual difference between the model and the data.

3.2 Spacetime near planets and black holes

The first full solution to Einstein’s field equations was derived by Karl Schwarzschild in 1915, and describes the behaviour of spacetime near a single concentration of mass. In other words, it describes the gravitational effects we see in everyday life, caused by the Earth’s gravity, as well as the effects of gravity close to other planets.

One of the solution's most interesting applications is to the spacetime around very dense objects: the **Schwarzschild metric** describes the existence of the objects now known as black holes. In this section you will examine the Schwarzschild metric and some of its peculiar consequences in the vicinity of very dense objects.

3.2.1 The Schwarzschild metric and its properties

The Schwarzschild metric is a ‘vacuum solution’ to the Einstein field equations. This means that it describes the geometry of regions that don’t themselves contain significant matter or energy (i.e. where the energy–momentum tensor is equal to zero). The geometry is instead determined by the presence of an external, nearby mass.

We can consider a region of spacetime within the gravitational influence of some sort of mass concentration, which could be a planet, a star or some other type of object, with total mass M . The influence of such a mass is spherically symmetric, and so the spatial part of the metric is usually written using spherical coordinates, r , θ and ϕ , whose origin is the centre of the dominating mass.

The Schwarzschild metric

$$\begin{aligned} ds^2 = & \left(1 - \frac{2GM}{c^2r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2r}\right)^{-1} dr^2 \\ & - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2 \end{aligned} \quad (3.2)$$

- What are the units of the metric coefficients for the ct and r coordinates?
- The term in brackets multiplying each coordinate is 1 minus a ratio of terms involving the central mass M and r . The numerator of this ratio has units of $G \times M$, which are $\text{m}^3 \text{s}^{-2}$, and the denominator has units of $c^2 \times r$, which are also $\text{m}^3 \text{s}^{-2}$. Hence the coefficients are dimensionless.

The $2GM/c^2r$ term in the ct and r coefficients can also be written as a dimensionless ratio of distances, R_S/r , where R_S corresponds to a radius known as the **Schwarzschild radius**:

$$R_S = \frac{2GM}{c^2} \quad (3.3)$$

The significance of this radius is that when the r coordinate is equal to R_S , the $c^2 dt^2$ term in the metric vanishes and the dr^2 term tends to infinity, hence the spacetime interval ds itself becomes infinite. In other words, something peculiar happens to the behaviour of spacetime at the Schwarzschild radius. You will explore behaviour at the Schwarzschild radius in the next section, but in the following exercise you will first investigate the size of this radius for different astronomical objects.

Exercise 3.2

Calculate the Schwarzschild radii for the Earth ($M_{\oplus} = 5.97 \times 10^{24}$ kg), the Sun ($M_{\odot} = 1.99 \times 10^{30}$ kg) and a neutron star ($M_{\text{NS}} = 2 M_{\odot}$). Comment on how these values compare to the actual radii of these bodies (which are 6400 km, 7.0×10^8 m and 15 km, respectively).

The previous exercise points to the importance of the *density* of the massive object for the behaviour of spacetime in its vicinity: at high densities the Schwarzschild radius starts to approach the size of the object itself. In fact, we have ample evidence for the existence of astronomical objects that are sufficiently dense for their Schwarzschild radius to be larger than their physical size. Such objects are known as black holes.

Before we discuss the properties of black holes and what actually happens at the Schwarzschild radius in more detail, the following example considers how the metric behaves at the opposite extreme, namely, at locations a long way from the massive object.

Example 3.1

The limit in which $r \gg R_S$ corresponds to spacetime at a very large distance from the dominating mass M . Write a simplified expression for the Schwarzschild metric in this limit, and compare this to our Newtonian understanding of the effect of gravity at large distances from a large mass.

Solution

We first combine Equations 3.2 and 3.3 to write the Schwarzschild metric in terms of R_S :

$$ds^2 = \left(1 - \frac{R_S}{r}\right) c^2 dt^2 - \left(1 - \frac{R_S}{r}\right)^{-1} dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2$$

In the limit where $r \gg R_S$, we know that R_S/r tends to $1/\infty$, i.e. 0. This means that the two terms in brackets both simplify to 1. The metric therefore becomes:

$$ds^2 = c^2 dt^2 - dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2$$

Ignoring the ct coordinate for a moment, you should recognise the last three terms in the metric: this is just -1 times the metric of a flat three-dimensional space, as written in spherical coordinates. With the ct term included, it is the Minkowski spacetime metric (see Section 2.2.3), written in spherical coordinates.

The example has therefore shown that spacetime at very large distances from a mass concentration has a flat geometry. This is consistent with our Newtonian expectation that the effect of gravity will be very much reduced at large distances from a particular object.

3.2.2 Black holes and the event horizon

As mentioned briefly in the preceding section, **black holes** are objects whose density is sufficiently high for their Schwarzschild radius to be located outside the mass concentration itself. This means that there is a region within R_S into which matter can fall and be trapped, without possibility of escape. Such a radius is also known as an **event horizon**.

If an object of mass m is launched from the surface of a spherical body of mass M and radius R then, in order to escape from the gravitational influence of the body, the kinetic energy of the object must exceed its gravitational potential energy. This leads to the definition of an **escape speed**, v_{esc} , that is high enough to overcome gravity:

$$\frac{1}{2}mv_{\text{esc}}^2 = \frac{GMm}{R}, \quad \text{and so} \quad (3.4)$$

$$v_{\text{esc}} = \sqrt{\frac{2GM}{R}} \quad (3.5)$$

If we consider the escape speed for an object with radius $R < R_S$, as is the case for a black hole, then we find that $v_{\text{esc}} > c$. In other words, for the densest objects there is a radius from within which matter and light (if similarly affected by gravity) cannot escape, unless travelling faster than the speed of light. This radius corresponds to what we now refer to as the event horizon for a non-spinning, spherical body.

The idea of ‘dark stars’, so dense that neither light nor matter can escape from them, dates back to the eighteenth century. The concept was revived in the twentieth century as a result of the Schwarzschild solution, together with advances in stellar astrophysics and the observational discovery in 1967 of highly dense neutron stars (pulsars) by Jocelyn Bell and Anthony Hewish. Evidence has since accumulated for the existence of black holes throughout the Universe. Table 3.2 summarises some of the many forms of evidence for astronomical black holes, a number of which are depicted in Figures 3.7 and 3.8.

Table 3.2 A summary of evidence for the existence of black holes.

Observations	Evidence
Star orbits at the centre of the Milky Way	Detailed monitoring of the motion of stars at the centre of the Milky Way over several decades demonstrates the existence of a concentration of invisible mass of $\sim 4 \times 10^6 M_\odot$ in a central, Solar System-sized region. (Figure 3.7a)
Gas orbits in galaxy centres	Doppler shift measurements of gas close to the centres of nearby galaxies show very high speeds, which require the presence of a central black hole.
X-ray observations of active galaxies	The extreme luminosities of active galaxies and quasars can only be explained by large amounts of matter falling in towards a black hole.
Imaging black-hole ‘shadows’	The Event Horizon Telescope has imaged the shadow of a black hole in the nearby galaxy M87: a dark central region surrounded by the light of infalling and outflowing matter. (Figure 3.7b)
Radio jets	The existence of powerful radio jets travelling at close to the speed of light from galaxy centres can only be powered by matter falling onto a central supermassive black hole. (Figure 3.7c)
Gravitational waves	Recent detections of gravitational waves provide evidence for many mergers of black holes produced by stellar evolution. (Figure 3.8)

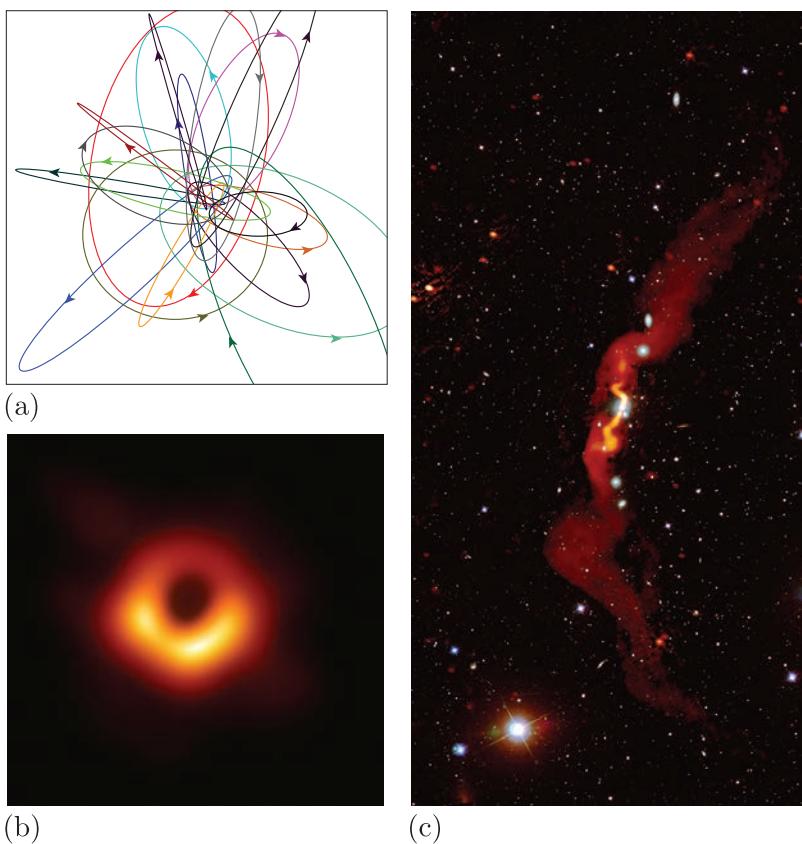


Figure 3.7 Observational evidence for black holes: (a) the orbits of stars around the centre of the Milky Way; (b) the Event Horizon Telescope image of a black-hole shadow; (c) powerful radio-emitting jets, travelling at close to the speed of light, powered by matter falling in towards a supermassive black hole in the central galaxy.

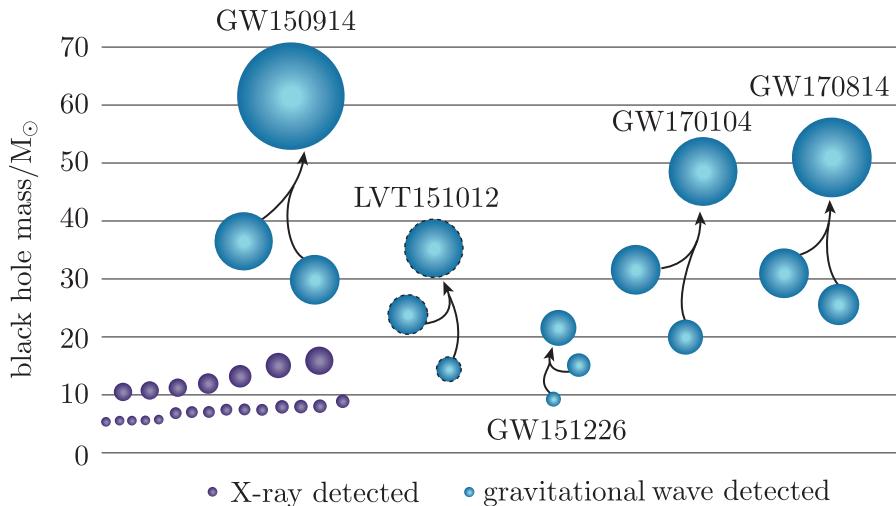


Figure 3.8 A chart of black-hole mergers (blue circles) discovered via gravitational waves, together with measured masses of some X-ray-detected black holes (purple circles). LVT151012 is a ‘candidate event’ that was too weak to be conclusively claimed as a detection.

3.2.3 Behaviour of spacetime near black holes

It is possible for individual particles (or large objects, e.g. an entire star or planet) to cross over an event horizon from the outside, but not to escape it from the inside. A body released from rest at a large distance from a non-rotating black hole requires only a finite proper time to reach the gravitational **singularity** at its centre (see Figure 3.9). At this location the metric tends to infinity, irrespective of the coordinate system used to describe it.

gravitational singularity

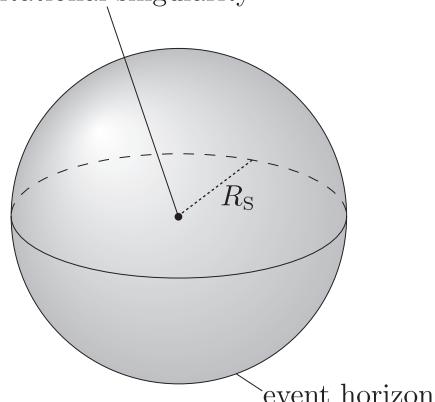


Figure 3.9 Simplified geometry of a black hole

What happens to mass that arrives at the singularity is the subject of theoretical speculation but, unlike behaviour near the event horizon, modern science provides no way to study it. However, it can be shown that the time taken for a body following a direct radial path (i.e. with constant θ and ϕ coordinates) to travel from the event horizon to the central singularity is

$$d\tau_{\text{sing}} = \frac{2}{3} \frac{R_S}{c} \quad (3.6)$$

The next exercise allows you to explore the timescales involved.

Exercise 3.3

- (a) What is the proper time interval required for a falling body to travel from the Schwarzschild radius to the singularity of a black hole with three times the mass of the Sun?
- (b) What is the corresponding proper time if instead we consider a supermassive black hole of mass $10^9 M_\odot$?

The previous exercise illustrated that matter can very quickly fall from outside a black hole to its very centre. An observer travelling with the infalling matter will not notice anything unusual happen when crossing the

event horizon. However, the situation looks very different to a distant observer watching matter approach a black hole.

We saw in Example 3.1 that for a distant observer (i.e. $r \gg R_S$) the metric is locally that of special relativity (the flat Minkowski metric). The next example explores what will be measured from this distant observer's perspective.

Example 3.2

A distant observer is located on the same radial line as an object falling into a black hole.

- Find an expression for the time interval $t_2 - t_1$ between the emission of a photon by the falling object near to the black hole at coordinate r_1 and the photon's arrival at the distant observer's location r_2 . You may assume that the relevant light signals are also travelling only in the radial direction, so that $d\theta$ and $d\phi$ can be neglected.
- Comment on how the time taken for light signals from near the event horizon to reach r_2 compares to the r coordinate separation divided by the speed of light, and explain what happens to the interval $t_2 - t_1$ when the location r_1 of the signal being emitted tends to R_S .

Solution

- The world line of a photon must be a null geodesic (see Section 2.3.3), and so the spacetime separation ds between the two events must be zero. Therefore, the Schwarzschild metric becomes

$$ds^2 = (1 - R_S/r) c^2 dt^2 - \frac{dr^2}{1 - R_S/r} = 0$$

Rearranging this expression and taking square roots gives

$$c dt = \frac{1}{1 - R_S/r} dr$$

We can integrate each side for a photon travelling from r_1 to r_2 :

$$c \int_{t_1}^{t_2} dt = \int_{r_1}^{r_2} \frac{1}{1 - R_S/r} dr$$

Dividing by c and evaluating the (simpler) left-hand integral gives:

$$t_2 - t_1 = \frac{1}{c} \int_{r_1}^{r_2} \frac{1}{1 - R_S/r} dr$$

The remaining integral can be solved by multiplying the top and bottom of the fraction by r and using the substitution $u = r - R_S$, so that:

$$t_2 - t_1 = \frac{1}{c} \int_{r_1 - R_S}^{r_2 - R_S} \frac{u + R_S}{u} du = \frac{1}{c} \int_{r_1 - R_S}^{r_2 - R_S} \left(1 + \frac{R_S}{u} \right) du$$

and so

$$t_2 - t_1 = \frac{r_2 - r_1}{c} + \frac{R_S}{c} \ln\left(\frac{r_2 - R_S}{r_1 - R_S}\right) \quad (3.7)$$

- (b) The logarithm term in Equation 3.7 means that the time interval must be larger than the radial separation divided by the speed of light. In the limit $r_1 \rightarrow R_S$, the denominator of the log term in Equation 3.7 tends to zero, so that the fraction inside the logarithm tends to infinity. Hence the time interval for a light signal to reach a distant observer from very close to the event horizon is infinite.

The calculation above reinforces one of the peculiar properties of black holes: that matter can fall into a black hole quite quickly, without anything unusual happening to it when passing the event horizon, but, to a distant observer, the matter is seen to get closer and closer to the event horizon, but not observed to cross it.

Figure 3.10a illustrates the path of infalling matter as viewed by an observer travelling with it (the proper time measured along the world line of the infalling matter), with panel (b) comparing this with what is seen by a distant observer (the coordinate time at which light signals from the infalling material at a particular radius will be received).

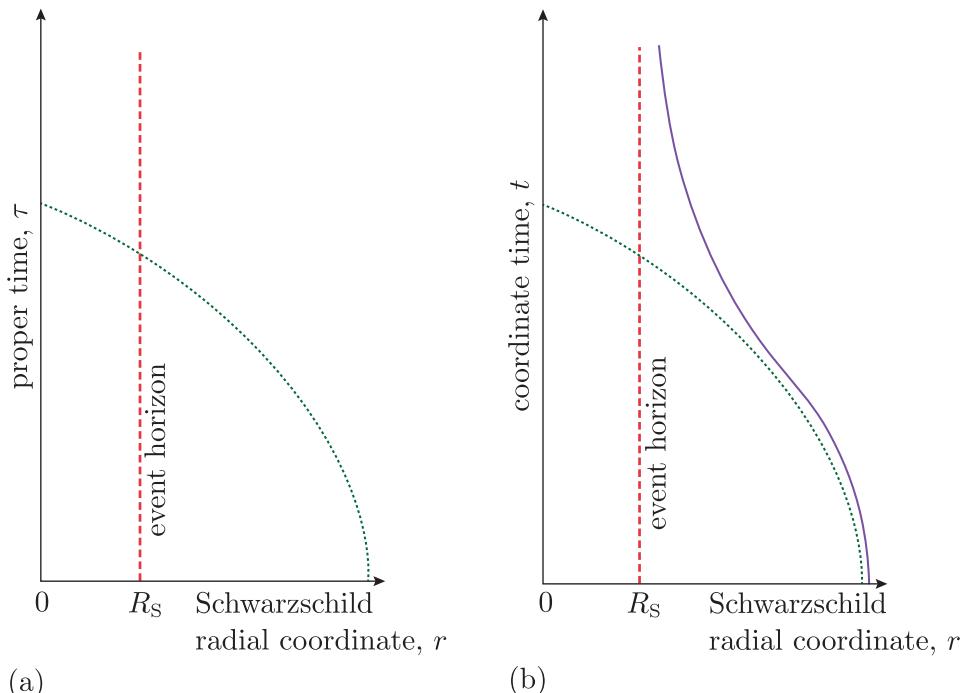
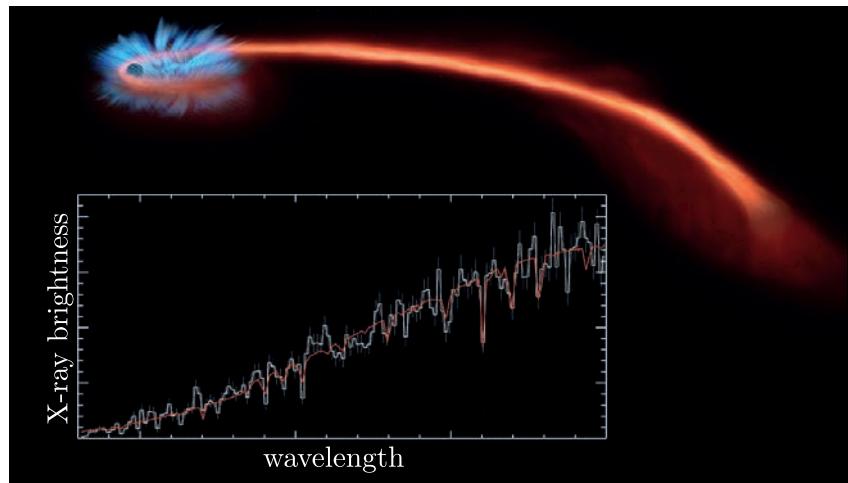


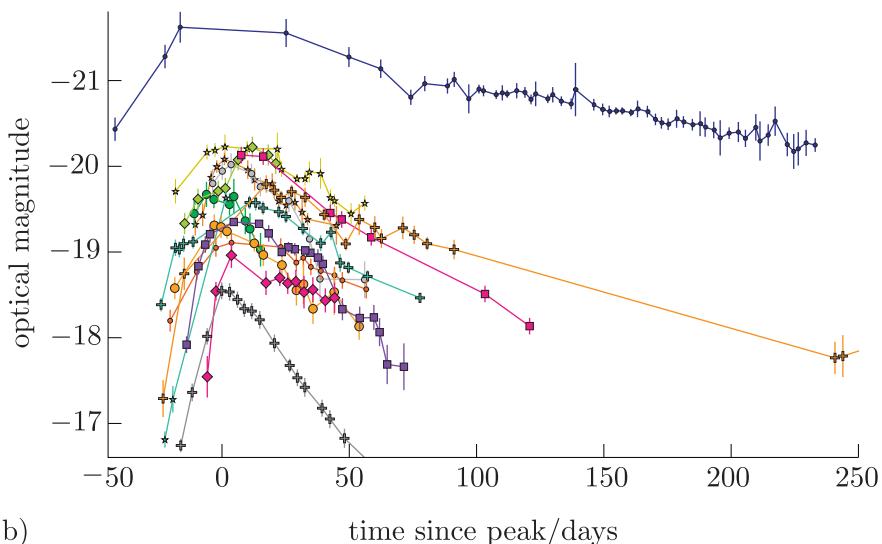
Figure 3.10 The path of an object falling radially into a black hole in Schwarzschild coordinates, as observed by: (a) an observer falling with the object (green dotted line indicates the proper time, τ , along the world line); (b) a distant observer at a location where the effects of the central mass are negligible (blue solid line indicating the coordinate time at which signals are received).

Of course, matter falling into a black hole will not always take a radial path, and it *will* also be affected by the strong gravitational field of the object. Tidal forces, that is, strong differences in the gravitational field strength at different nearby locations, can have a very dramatic effect on an infalling body, depending on the black hole's mass.

Figure 3.11 shows an artist's impression of such tidal disruption events (TDEs), as well as some observational evidence for them. The plots in panels (a) and (b) show X-ray and optical flares respectively, which are thought to occur as stars fall into black holes. Such observations can allow astronomers to determine the properties of the star being disrupted.



(a)



(b)

Figure 3.11 Observational evidence of stars disrupted by tidal forces in the vicinity of black holes: (a) an artist's impression of a TDE, with an X-ray spectrum showing peaks in brightness as more material falls into the black hole; (b) optical light curves from a sample of TDEs, showing how they brighten and then fade with time (after van Velzen *et al.*, 2021).

The discussion above has demonstrated that gravity can cause time dilation (disagreement about the time interval between events), just as time dilation can be caused by motion at relativistic speeds. Another consequence of gravitational time dilation is a shifting in the frequency of light signals, because the frequency of light is the inverse of the time interval between successive wavefronts.

The decrease in frequency (increase in wavelength) of an electromagnetic signal emitted close to a massive object when observed at a distance that is less influenced by the object is termed **gravitational redshift**, and is given by Equation 3.8.

Gravitational redshift

For an electromagnetic signal emitted at a distance r_{em} from an object of mass M , the observed frequency at large distance, ν_∞ , is related to the emitted frequency, ν_{em} , by

$$\nu_\infty = \nu_{\text{em}} \left(1 - \frac{2GM}{c^2 r_{\text{em}}}\right)^{1/2} \quad (3.8)$$

As noted in Section 3.1.3, gravitational redshift does not require a black hole – measurements of the Earth’s gravitational redshift are one of the ways general relativity has been experimentally verified – but this effect is a crucial one for interpreting astronomical observations of black holes.

To close, we have only been able to provide a brief introduction to the spacetime of black holes in this chapter. An important complication not discussed here is that the Schwarzschild metric applies to *non-rotating* black holes. In reality, we expect that the majority of black holes are spinning (indeed, it is possible to make measurements of their spin in some cases), which alters the metric. The metric for a spinning black hole is the **Kerr metric**, which has both inner and outer event horizons, and some further interesting behaviour near to those regions. We will return to the topic of black holes in the second part of the module.

3.3 The geometry of the expanding Universe

The previous section introduced one important example of a curved spacetime metric, in which a single concentration of mass influences the geodesics for matter and light and how they are seen by different observers. In this section we consider how to construct a metric appropriate for describing the geometry and time evolution of the entire Universe. This is a rather ambitious undertaking, but the cosmological principle – which you met in Chapter 1 – helps make it possible to construct a metric straightforward enough to use to build models of the Universe that can be directly tested against astronomical observations in many different ways.

3.3.1 Cosmic time and co-moving coordinates

The starting point for constructing a metric for the entire Universe and its expansion must be the behaviour we infer from astronomical observations made from our cosmic location on Earth. However, the previous section has shown that observers may differ in the conclusions they draw about events, either because they are in relative motion or because they are at different locations relative to concentrations of mass.

We cannot travel to distances far enough away on cosmic scales to get a different perspective, so it's only possible to proceed by making some big assumptions about the nature of the Universe and then finding ways to test whether all of the available evidence supports them. The first necessary assumption is the cosmological principle: on large scales, the Universe is homogeneous and isotropic.

- What evidence do we have that the cosmological principle is correct?
- Two key observations are that maps of the large-scale structure of the Universe (clusters, superclusters and voids between them) appear homogeneous and isotropic on 100 Mpc scales, and the cosmic microwave background (corrected for distortion due to local motions) is uniform in all directions, with fluctuations of less than 1 part in 10 000 (see Chapter 1).

The cosmological principle makes no mention of time, but we know from observations that the large-scale structure of the Universe is *not* constant in time – it is expanding. A second crucial assumption, known as **Weyl's postulate** after its originator Hermann Weyl, greatly simplifies the way time is incorporated into cosmological models. It proposes a set of hypothetical observers whose motion in spacetime matches the overall expansion of the Universe, sometimes known as the **Hubble flow**. These **fundamental observers** will perceive the Universe as obeying the cosmological principle, and will agree about the nature of the Universe's expansion.

Weyl's postulate

In cosmic spacetime there exists a set of fundamental observers, whose world lines are time-like geodesics that do not ever meet, except possibly at an initial singularity in the past and/or a final singularity in the future.

- Where might such fundamental observers be located?
- A hypothetical fundamental observer could be moving with their local supercluster or their home galaxy. In the latter case they would need to correct for any local motions within their wider environment, such as a galaxy orbit within a cluster or supercluster.

Figure 3.12 illustrates how the existence of fundamental observers (which could, for example, represent the locations of particular galaxies moving with the Hubble flow) allows a universal definition of **cosmic time** to be agreed. Hypersurfaces are shown to represent the spatial dimensions of the Universe at a particular time. The figure uses a two-dimensional surface to represent the spatial dimensions of the Universe, but in reality expansion is taking place in three dimensions. This is why the surfaces are labelled as ‘hypersurfaces’, indicating they have more than two spatial dimensions.

The spatial separations between the fundamental observers, as measured on the hypersurfaces, increase with time. But the cosmological principle requires that time elapses at the same rate along the world lines of all fundamental observers, so that all such observers can agree on the rate at which the spatial dimensions of the Universe expand.

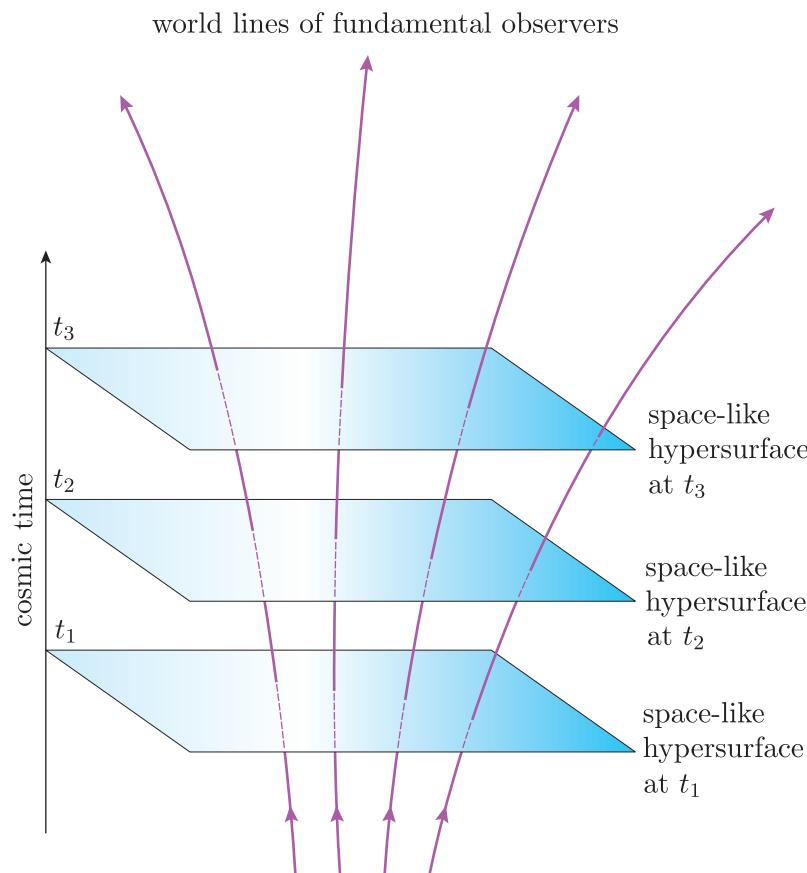


Figure 3.12 The world lines in cosmic spacetime of fundamental observers who see the Universe as homogeneous and isotropic. The vertical axis represents a universal cosmic time that the fundamental observers can agree on, while the 2D surfaces represent the 3D spatial geometry that expands homogeneously and isotropically with time.

Having defined a cosmic time in relation to the Hubble flow, it is now possible to determine a useful set of spatial coordinates that describe the hypersurfaces of Figure 3.12. We can specify a grid of coordinates such that the world line of a particular fundamental observer is assigned the

same coordinate values at all times, that is, for every instance of the space-like hypersurface in Figure 3.12 (corresponding to a particular cosmic time). These values are known as **co-moving coordinates**, and although the grid on which they're based expands with cosmic time, the coordinate distances between locations stay fixed, as illustrated in Figure 3.13.

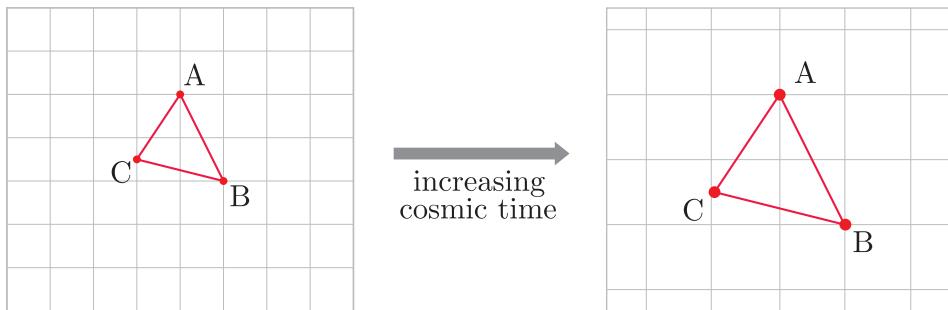


Figure 3.13 A grid of co-moving coordinates will track the expanding Hubble flow. Points A, B and C move with the flow and so have coordinate values that remain constant with cosmic time.

- The separation distances between fundamental observers are now twice as large as they were at a particular time in the past, $t_{1/2}$. How does the current co-moving distance between the Milky Way and a distant galaxy – NGC 1262, for example – compare to their co-moving distance at time $t_{1/2}$?
- The co-moving distance between the Milky Way and NGC 1262 will be unchanged (except from any correction for local motions that are not part of the Hubble flow).

The co-moving coordinate distance at a particular time is not a useful measurement for all possible purposes. For example, it won't, on its own, tell us how long it would take for light or a spacecraft to travel between two galaxies. But co-moving coordinates provide a straightforward way to simplify the metric for cosmic spacetime.

The final ingredient we need for a cosmological metric is the **scale factor**, $a(t)$, where t is the cosmic time, which describes the relationship between co-moving coordinate distances and the true physical distance between objects at a particular cosmic time. In other words, the scale factor captures the rate at which the coordinate grid in Figure 3.13 expands.

- Which parameter that describes the Universe's expansion and that you met earlier in the module must be closely related to the scale factor $a(t)$?
- The Hubble constant, introduced in Chapter 1, describes the rate of expansion of the Universe at the present time, and so must be related to the scale factor. (The precise relationship will be discussed in a later section.)

Together, the concepts of a universal cosmic time, co-moving coordinates, and a scale factor to describe cosmic expansion allow a metric to be constructed to describe the large-scale geometry of the Universe. The next section introduces that metric.

3.3.2 The Robertson–Walker metric

To get a feel for the form that a cosmological metric must take, we will initially make the simplifying assumption (which is not necessarily correct) that the large-scale spatial geometry of the Universe is flat. That assumption leads to the following metric for an expanding spacetime:

$$ds^2 = c^2 dt^2 - a^2(t) (dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2) \quad (3.9)$$

where t is cosmic time, $a(t)$ is the scale factor, and r , θ and ϕ are co-moving coordinates for a three-dimensional spherical geometry centred on a fundamental observer.

- How does the metric in Equation 3.9 differ from the Minkowski metric of the previous chapter, and why?
- The spatial part of the metric shown in Equation 3.9 is expressed in spherical coordinates, but the part inside the brackets corresponds to flat space (see Section 2.3.1). This spatial part is then multiplied by a scaling factor a , which depends on time. This means that although the geometry changes with time, at any particular cosmic time, t , it is equivalent to the Minkowski metric.

The question of whether the Universe is spatially flat (in which case Equation 3.9 correctly describes its geometry) or curved (so that a more complex metric is needed) is one of the most important in cosmology. You will see later that this question has significant implications for the eventual fate of the Universe.

- What physical properties of the Universe would you expect to determine whether or not it has spatial curvature?
- Einstein's field equations relate curvature to the overall mass and energy content of the Universe. Therefore, the amount of matter and energy in the Universe should determine its curvature.

You will notice that, in contrast to the Schwarzschild metric (Equation 3.2), there is no obvious mass term in Equation 3.9, and so it is not immediately clear that it can be a solution to Einstein's field equations. The explanation for this is that the scale-factor term, $a(t)$, depends on the mass and energy content of the Universe, which influence how the scale factor changes with time. We will explore how a is related to mass and energy in the next chapter, but it is sufficient to note that this relationship leads us to cosmological models that we can test directly with observations.

The best observational evidence we have at the moment indicates that the Universe *is* spatially flat, as you will see later. But we need to incorporate the possibility of curvature into the metric to be able to explore the range of possible cosmological models and understand how modern observations are constraining them.

The full cosmological metric, including curvature of space, is known as the **Robertson–Walker metric**. It is most commonly written as follows, where k is a ***spatial curvature parameter***:

The Robertson–Walker metric

$$ds^2 = c^2 dt^2 - a^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right) \quad (3.10)$$

For a flat 3D geometry, $k = 0$, but it can also take positive or negative values,* indicating positively and negatively curved geometries, respectively (as introduced in Section 2.3.1). For the $k = 0$ case, Equation 3.10 simplifies to Equation 3.9.

It is also important to note that because the 3D spatial geometry (the part of the Robertson–Walker metric inside the brackets) is time-independent (as illustrated in the 2D schematic of Figure 3.12), k is also constant with time. It does not describe the curvature of four-dimensional spacetime, but only the curvature of the co-moving grid that scales uniformly with $a(t)$.

- What are the units for the quantities $a(t)$ and k ?
- Each term in the metric must have the same units as ds^2 , which has units of distance squared (e.g. m^2). This requires the scale factor a to be dimensionless, and k to have units of the inverse of distance squared (e.g. m^{-2}).

3.3.3 The Hubble parameter and the scale factor

We conclude this chapter by making some initial connections between the rather abstract topic of metrics and the quantities that can be directly measured using astronomical observations.

A crucial observation is that the Universe is expanding at the present time; you saw in Chapter 1 that this expansion is measured using a quantity known as the Hubble constant, H_0 , which corresponds to the rate of expansion over a given length scale. H_0 relates to the cosmic scale factor a according to

$$H_0 = \frac{1}{a(t_0)} \frac{da}{dt}(t_0) = \frac{\dot{a}(t_0)}{a(t_0)} \quad (3.11)$$

where t_0 is the present time. We have also introduced the standard notation $\dot{a} = da/dt$; throughout the module, a dotted quantity corresponds to the time derivative of that quantity.

The next exercise involves considering quantitatively how the scale factor has changed with time over the recent history of the Universe.

*Some cosmology textbooks use a scaled coordinate system in which k is normalised to take only one of the three values $+1$, 0 or -1 , rather than allowing any value for k as we do here. This requires a slightly different definition for the scale factor. Either convention makes certain cosmological calculations simpler, but for the purposes of this book k is not normalised and can take any value.

Exercise 3.4

Assuming (for the purpose of this exercise) that $H_0 = 68 \text{ km s}^{-1} \text{ Mpc}^{-1}$, and has not changed significantly over the past billion years, calculate the factor by which the scale factor of the Universe has changed:

- (a) over the last 1000 years
- (b) over the last 100 million years.

(Hint: assume the scale factor at the present day is $a(t_0) = 1$.)

Measuring H_0 can therefore give us information about how a is changing. But the assumption made in Exercise 3.4 that H_0 is the same for all cosmic epochs is not correct: in some cosmological models it can change significantly over the lifetime of the Universe. A more general **Hubble parameter** is defined as

$$H(t) = \frac{\dot{a}}{a} \quad (3.12)$$

where H now represents the value of the Hubble constant at some particular time t , which could be in the past or the future.

The scale factor is also related to another important observational quantity linked to cosmic time: the cosmological redshift. The redshift of a distant light signal (z) is related to the emitted and observed wavelengths (λ) or frequencies (ν) of the signal according to

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}} \quad (3.13)$$

We can derive a relationship between a and z by considering the behaviour of a light signal from a distant galaxy, located at a fixed radial coordinate, r_{gal} . The light signal is emitted at time t_{em} and observed at time t_{obs} .

To simplify the algebra, we will start by using the Robertson–Walker metric for a flat geometry ($k = 0$) only. Recalling that light follows null geodesics, for which $ds^2 = 0$, the metric can be greatly simplified in this situation. The light is also travelling on a purely radial path, so the $d\theta$ and $d\phi$ terms are zero because we can define the distant galaxy to have the same θ and ϕ coordinates as the observer. Equation 3.10 therefore becomes

$$0 = c^2 dt^2 - a^2(t) dr^2$$

We can obtain an expression for the radial coordinate of the galaxy, r_{gal} , by first rearranging and then taking square roots to obtain

$$\frac{c dt}{a(t)} = dr$$

and we can then integrate to get

$$c \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)} = \int_0^{r_{\text{gal}}} dr = r_{\text{gal}}$$

Here, the integration limits for the left-hand integral correspond to the time range between the light being emitted and observed, and those for the right-hand integral correspond to the change in coordinate distance for the same spacetime interval under consideration.

How can we use this to find a relationship with redshift? Because z is related to wavelength, which can be thought of as the interval between the peaks of light signals, we can start by considering a second signal, corresponding to the next wavefront of the light from the galaxy. This signal is emitted at time $t_{\text{em}} + \Delta t_{\text{em}}$ and observed at a time $t_{\text{obs}} + \Delta t_{\text{obs}}$.

The situation here is the same as for the first signal, except that the start and end times for the left-hand integral are modified:

$$c \int_{t_{\text{em}} + \Delta t_{\text{em}}}^{t_{\text{obs}} + \Delta t_{\text{obs}}} \frac{dt}{a(t)} = \int_0^{r_{\text{gal}}} dr = r_{\text{gal}}$$

Because the coordinate distance travelled by the light, r_{gal} , will be the same in both cases, the two expressions can be equated:

$$c \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)} = c \int_{t_{\text{em}} + \Delta t_{\text{em}}}^{t_{\text{obs}} + \Delta t_{\text{obs}}} \frac{dt}{a(t)}$$

We don't actually know how a depends on time, and so we need to find a way to take it out of the integral. Although the scale factor may change significantly between t_{em} and t_{obs} – that is, $a(t_{\text{em}}) \neq a(t_{\text{obs}})$ – it cannot change significantly within the timescale between the emission of the two wavefronts or that of their receipt (e.g. see the solution to Exercise 3.4). Therefore, we can take $a(t_{\text{em}}) = a(t_{\text{em}} + \Delta t_{\text{em}})$ and $a(t_{\text{obs}}) = a(t_{\text{obs}} + \Delta t_{\text{obs}})$.

In order to make use of these simplifications we need to take the (not at all obvious) step of subtracting the integral

$$c \int_{t_{\text{em}} + \Delta t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)}$$

from both sides, which effectively changes the integration ranges to give

$$\int_{t_{\text{em}}}^{t_{\text{em}} + \Delta t_{\text{em}}} \frac{dt}{a(t)} = \int_{t_{\text{obs}}}^{t_{\text{obs}} + \Delta t_{\text{obs}}} \frac{dt}{a(t)}$$

where we have also cancelled out factors of c . Now both integrals only cover a very small time interval (e.g. for typical light frequencies the interval between wavelengths is $\sim 10^{-15}$ s) compared to the scale on which the Universe expands, and so now we can take a to be constant for each integral, giving

$$\frac{1}{a(t_{\text{em}})} \int_{t_{\text{em}}}^{t_{\text{em}} + \Delta t_{\text{em}}} dt = \frac{1}{a(t_{\text{obs}})} \int_{t_{\text{obs}}}^{t_{\text{obs}} + \Delta t_{\text{obs}}} dt$$

Integrating now gives us

$$\frac{\Delta t_{\text{em}}}{a(t_{\text{em}})} = \frac{\Delta t_{\text{obs}}}{a(t_{\text{obs}})} \quad (3.14)$$

In the following exercise you will use this expression to finally link a and redshift.

Exercise 3.5

Noting that the frequency of light is the inverse of the time interval between successive wavefronts (i.e. $\Delta t = 1/\nu$), use Equation 3.14 to find an expression relating scale factor and redshift, i.e. a and z .

For simplicity, the discussion and exercise above considered only the case of $k = 0$; however, the same derivation can be carried out for the full Robertson–Walker metric. The only difference is that the term in r_{gal} is more complicated, but it still corresponds to the same distance measure for both wavefronts. Hence the same relationship between a and z holds for all geometries.

Scale factor and redshift

$$a \propto \frac{1}{1+z} \quad (3.15)$$

The redshift of a distant galaxy therefore tells us how much smaller the scale factor of the Universe was at the time the light we measure was emitted.

Conventionally, the present-day value of the scale factor, $a(t_0)$, is set to 1, and so Equation 3.15 becomes an equality.

The final exercise in this chapter allows you to further explore the relationship between redshift and the scale factor.

Exercise 3.6

Two distant galaxies are measured to have redshifts of $z = 10.1$ and $z = 3.6$. Calculate by what factor the distance between two locations in the Universe has increased in size since the time when the light we observe from each galaxy was emitted.

3.4 Summary of Chapter 3

- The (weak) **equivalence principle** states that within a local region of spacetime, close to a concentration of mass, the motion of an object cannot be distinguished by any experiment from how it would behave in a region of (appropriately chosen) uniform acceleration.
- The theory of general relativity is a metric theory of gravity, encapsulated in **Einstein's field equations** (Equation 3.1). These equations relate a function of the curvature tensor to the **energy-momentum tensor**, such that ‘matter tells space[time] how to curve’ and ‘space[time] tells matter how to move’ (Misner, Thorne and Wheeler, 1973, p. 5).
- The predictions of general relativity are well tested, and supported by a range of evidence including measurements of the **precession** of the perihelion of Mercury, deflection of light passing near to the Sun, and the measurements of **gravitational redshift** of the **Pound–Rebka experiment**. **Shapiro delay** and both indirect and direct detection of **gravitational waves** also support the predictions of the theory.
- Einstein's field equations are effectively a set of 16 separate equations that can, in principle, be solved for a given energy-momentum tensor (mass and energy distribution) to determine the metric.
- The **Schwarzschild metric** is a solution to the Einstein field equations that applies to regions in the vicinity of a localised mass, and so describes the behaviour of spacetime near astronomical objects such as planets, stars and **black holes**:

$$ds^2 = \left(1 - \frac{2GM}{c^2r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2r}\right)^{-1} dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2 \quad (\text{Eqn 3.2})$$

- In the vicinity of a black hole the Schwarzschild metric predicts the presence of an **event horizon**, defined in the case of a non-rotating black hole by the **Schwarzschild radius**, from within which no light signals nor matter can exit. Distant observers measure time intervals for events close to the event horizon as tending towards becoming infinite, and measure light signals emitted near the central mass to be gravitationally redshifted.
- A metric to describe the geometry and expansion of the Universe can be constructed by defining a set of **fundamental observers** whose locations trace the **Hubble flow** and define a set of fixed **co-moving coordinates**. All such observers agree on a definition of **cosmic time**.
- The resulting metric is known as the **Robertson–Walker metric** and describes the spacetime of a homogeneous and isotropic expanding Universe:

$$ds^2 = c^2 dt^2 - a^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right) \quad (\text{Eqn 3.10})$$

- The **scale factor**, a , describes the rate at which any region of the Universe expands (or contracts) with time and is related to cosmic redshift, z , via

$$a \propto \frac{1}{1+z} \quad (\text{Eqn 3.15})$$

- The **Hubble parameter**, $H(t)$, describes the rate at which the scale factor evolves with time over a given length scale, and is defined as

$$H(t) = \frac{1}{a} \frac{da}{dt} = \frac{\dot{a}}{a} \quad (\text{Eqn 3.12})$$

The value of the Hubble parameter in the present epoch is known as the Hubble constant, H_0 .

- The spatial curvature of the Universe is parameterised by the (time-independent) **curvature parameter**, k , which is zero for a flat geometry, but can also take positive or negative values, indicating curved geometries.

Chapter 4 Cosmological models and their key parameters

The first three chapters of this book have introduced fundamental physical concepts including the cosmological principle and the theories of special and general relativity, and set out the metric for an expanding Universe and its key parameters.

In this chapter you will learn how those concepts can be used to build mathematical models that describe the Universe and its contents. Using these models it becomes possible to understand how the Universe has evolved in the past, and predict how it will continue to evolve in the future.

Objectives

Working through this chapter will enable you to:

- derive the two Friedmann equations and the fluid equation, and understand how they can be used to describe and model the expansion of the Universe
- use the Friedmann equations and the fluid equation to model the behaviour of some simple idealised universes
- predict how the densities and expansion rates of model universes containing different proportions of matter and radiation evolve with time
- understand where the cosmological constant appears in the Friedmann equations and the fluid equation, and explain how its existence affects the evolution of the Universe
- describe how the geometry and contents of the Universe influence its history and its ultimate fate.

4.1 The Friedmann equations

This section introduces two equations called the Friedmann equation and the acceleration equation, which are often collectively referred to as the Friedmann equations. Together with a third equation called the fluid equation, which we will also consider, they can be used to describe how the expansion, geometry and contents of the Universe depend on one another.

4.1.1 The Friedmann equation

The **Friedmann equation** is one of the most important equations in cosmology. It describes how the expansion rate of the Universe depends on its geometry, and the density of matter and energy within it.

A completely rigorous derivation of the Friedmann equation would entail solving Einstein's field equations and assuming that spacetime is described by the Robertson–Walker metric. However, such a derivation requires a much more mathematical foundation in general relativity than this module

provides. For that reason, we will now derive the Friedmann equation using an approach that assumes that spacetime is Euclidean and uses arguments from Newtonian dynamics, and has a more intuitive physical description. We will still arrive at an expression that is equivalent to the general relativistic one we would have derived otherwise.

We start by considering the cube depicted in Figure 4.1, which shows part of a hypothetical universe. It is expanding *isotropically*, and is filled with a completely smooth and *homogeneous* fluid that has density $\rho(t)$ at time t . We will assume that the total fluid content of the universe is constant, so $\rho(t)$ decreases as the universe expands.

We can pick a point P in this hypothetical universe and imagine a spherical region with radius R , which is centred on P and is expanding along with the universe. A small test particle with mass m lies on the boundary of the sphere, and M_{sphere} is the total mass of material inside the spherical region.

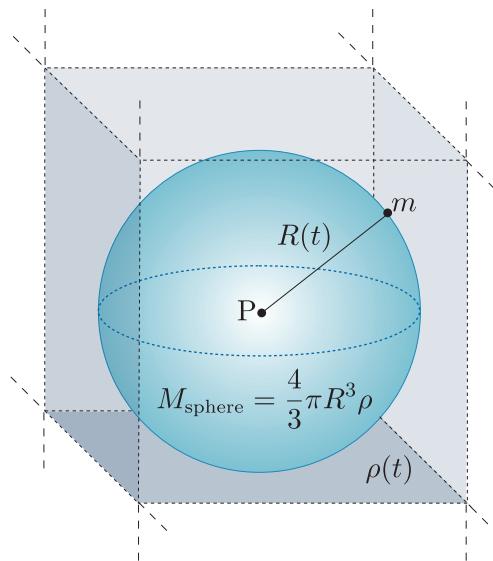


Figure 4.1 Part of a hypothetical universe filled with a homogeneous medium that has density $\rho(t)$ at time t . A spherical region with radius $R(t)$ is centred on a point P , with a small test particle of mass m on its boundary.

- Does it matter which point within the universe we choose to label as P ?
- No. Our assumption of homogeneity means that any point in our hypothetical universe is completely identical to all others. Similarly, all other spheres with radius R centred on these arbitrary points are also completely equivalent.

Let's compute the total energy U of the test particle on the boundary of the spherical region. The particle will experience a gravitational force in the direction of P , given by

$$F = \frac{GM_{\text{sphere}}m}{R^2}$$

where $M_{\text{sphere}} = (4/3)\pi R^3 \rho(t)$ is the total mass of the material within the

spherical region. This means that the gravitational potential energy of the particle is

$$E_g = -\frac{GM_{\text{sphere}}m}{R} = -\frac{4}{3}G\pi R^2 \rho m$$

What about all the material in the universe that sits outside the sphere we've defined: does it also exert a gravitational force on the test particle? In fact, the answer is no. Newton was the first to show that a spherical shell of matter does not exert a net gravitational force on any object inside it. This result is true regardless of the object's location within the shell. To see that only the material inside the sphere exerts a gravitational force on our test particle, we could just divide all the remaining fluid in the universe (i.e. external to the sphere) into a series of concentric spherical shells, none of which would exert any force on our particle.

The only force acting on the particle points back towards P along the radial direction, so its kinetic energy must just be

$$E_k = \frac{1}{2}m\dot{R}^2$$

where \dot{R} denotes the time derivative of position along the radial direction. Energy conservation implies that the total energy in the system is constant, so

$$U = E_k + E_g = \frac{1}{2}m\dot{R}^2 - \frac{4}{3}G\pi R^2 \rho m = \text{constant} \quad (4.1)$$

Our assumption of a homogeneous, isotropically expanding universe means that we can relate the physical distance $\mathbf{R}(t)$ between *any* two points at time t to the co-moving distance \mathbf{x} between them (see Section 3.3.1) using

$$\mathbf{R}(t) = a(t)\mathbf{x}$$

where $a(t)$ is the scale factor of the universe at time t , and \mathbf{x} is time-invariant by definition. We can now rewrite Equation 4.1 in terms of a and $x \equiv |\mathbf{x}|$:

$$U = \frac{1}{2}m\dot{a}^2 x^2 - \frac{4}{3}G\pi a^2 x^2 \rho m \quad (4.2)$$

- If \dot{a} is positive, is the universe expanding or contracting?
- If $\dot{a} > 0$ then the scale factor is increasing with time and the universe must be expanding.

Rearranging Equation 4.2 and defining a quantity $k = -2U/(mc^2 x^2)$ yields our Newtonian expression for the Friedmann equation.

The Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} \quad (4.3)$$

Recall from the previous chapter that 'dot notation' can be used to indicate time-derivatives of quantities, so \dot{x} means dx/dt and \ddot{x} means d^2x/dt^2 .

Before comparing this expression with its equivalent general-relativistic form, we can observe some of its general properties. Note that changing the sign of \dot{a} in Equation 4.3 does not change the value of the term on the left-hand side, which only depends on \dot{a}^2 . Indeed, there is nothing in our derivation that formally requires $\dot{a} > 0$, which means that the Friedmann equation is equally valid when considering contracting *or* expanding universes.

The parameter k in Equation 4.3 is the spatial curvature parameter that appears in the definition of the Robertson–Walker metric. The value of k does not depend on spatial coordinates *or* time – it is an invariant property of the Universe. The following example builds on our derivation so far, to show that this must be the case.

Example 4.1

Assuming that the Universe is homogeneous and undergoing isotropic expansion, explain why the Friedmann equation implies that k must be spatially invariant.

Solution

Consider the three terms in the Friedmann equation. In a homogeneous universe undergoing isotropic expansion, a , \dot{a} and ρ must be spatially invariant, so neither $(\dot{a}/a)^2$ nor $8\pi G\rho/3$ can depend on the value of x . Bearing in mind the spatial invariance of a , this means that the value of k must also be independent of x , and therefore spatially invariant, for the equality in Equation 4.3 to hold at all points in space.

The value of k has profound implications for the ultimate fate of universes that are described by the Friedmann equation. To see this, first note that the physical impossibility of negative densities means that the first term on the right-hand side of Equation 4.3 must always be positive.

Now, consider the case when $k < 0$, so both sides of Equation 4.3 are always positive. In an initially expanding universe, where $\dot{a} > 0$, the right-hand side of the equation must always be positive and so \dot{a} cannot change sign. Any expanding universe that has a negative value for k will continue to expand forever.

When $k > 0$ in a universe that is initially expanding, the scale factor will increase and ρ will decrease until

$$\frac{8\pi G\rho}{3} = \frac{kc^2}{a^2} \implies a^2 = \frac{3kc^2}{8\pi G\rho}$$

The arrow symbol \implies is sometimes used as shorthand for ‘implies’.

When this happens, \dot{a} will be zero and the universe will stop expanding. However, the second time derivative \ddot{a} will be negative, which means the universe will begin to contract again. Universes with positive k will eventually recollapse to zero size at an epoch known colloquially as the ‘big crunch’.

- How does a evolve in the limiting case when $k = 0$?
- When $k = 0$, the expansion continues forever but becomes slower and slower as \dot{a} and ρ both tend towards zero.

As the following highlight box explains, cosmologists often use shorthand names to refer to universes with different, specific values of k .

Open, closed and flat universes

Hypothetical universes that have negative values of k are often described as **open universes**, while those with positive k are often referred to as **closed universes**. The term **flat universe** is used to refer to the special case when k equals zero.

The nomenclature comes from the geometric interpretation of k representing the shape of spacetime in the context of the Robertson–Walker metric, which you learned about in Section 3.3.2.

We will return to these different types of universe later in the chapter when we consider further models, but before we end this section on the Friedmann equation let's explore how our Newtonian expression (Equation 4.3) differs from the relativistic one that can be derived by solving the Einstein field equations. The relativistic expression is

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2} \epsilon - \frac{kc^2}{a^2} \quad (4.4)$$

The only difference is that the density term ρ has been replaced by an energy density ϵ divided by c^2 . General relativity states that the *energy* of a group of particles is the relevant quantity to consider when determining their gravitational influence, not their rest mass. For a single particle with rest mass m_0 and momentum p , the relativistic expression for its energy is

$$E = \sqrt{p^2 c^2 + m_0^2 c^4} \quad (4.5)$$

This means that for a collection of massive particles with $v \approx c$, their *motion* contributes a significant fraction of their overall energy density. Another consequence of general relativity is that even massless particles such as photons exert a gravitational influence, because their energies contribute to the energy density of any volume they occupy.

Exercise 4.1

For the remainder of this chapter we will write the energy density explicitly as ρc^2 , rather than using the symbol ϵ that appeared in Equation 4.4.

With this in mind, and assuming a universe filled with non-relativistic particles that have velocities $v \ll c$, show that $\epsilon/c^2 \rightarrow \rho$, and therefore that the general-relativistic and Newtonian forms of the Friedmann equation are equivalent.

You should now have a good sense of the general properties of the Friedmann equation, and its capacity to predict how the geometry and density of the Universe affect its expansion. In the next section you will learn about another equation that describes the *inverse* relationship, and models how the densities of the Universe's contents depend on its expansion rate.

4.1.2 The fluid equation

In this section we will construct an equation called the **fluid equation** that describes how the energy densities, $\rho(t)c^2$, and pressures, P , of different fluids evolve in an expanding universe. Once we know how ρ changes with time, we will solve the Friedmann equation to predict the value of a at any time in the Universe's past or future.

Our derivation will consider the spherical region illustrated in Figure 4.1, and we will focus on the properties of the fluid it contains. Remember that the sphere is expanding isotropically along with the universe in which it is embedded, so its *co-moving* radius, $R' = R(t)/a(t)$, is constant. To simplify expressions in the rest of this section, we will assume that $R' = 1$.

We start with the first law of thermodynamics, which expresses how the heat Q flowing into and out of a fluid at pressure P is balanced by changes in its internal energy E and volume V .^{*} For the fluid in our model sphere, we can write the change in heat with respect to time as

$$\frac{dQ}{dt} = \frac{dE}{dt} + P \frac{dV}{dt} = 0 \quad (4.6)$$

- Suggest one reason why we can assume that $dQ/dt = 0$ for the fluid in our expanding spherical region.
- The simplest way to see that this is true is to realise that isotropic expansion preserves the homogeneity of the fluid. Since there can be no temperature gradients in a homogeneous medium, no heat can flow into or out of the volume.

Alternatively, we could recognise that the only way that heat can be transferred into or out of the region is if some of the fluid filling the universe crosses its boundary. We have assumed that this fluid is completely homogeneous, which means that there are no overdensities or pressure gradients that might induce flows within it. In other words, the expanding fluid remains completely static in the *co-moving* coordinate frame.

We have also specified that the spherical region has a constant co-moving radius. This means that no fluid enters or leaves the region because its boundary is completely fixed in the same coordinate frame in which the fluid remains static. It follows from this observation that dQ/dt must be equal to zero.

^{*}It is important to recognise that the pressure P of a fluid may, in fact, depend on other quantities, such as the fluid's density or the ambient temperature, that vary with time. For conciseness, we will not explicitly indicate the time-dependence of $P(t)$ in the remainder of this section, but we will return to examine it in Section 4.2.1.

To derive the fluid equation, we need to express dE/dt and dV/dt in terms of a and ρ and their time derivatives. The volume of the spherical region in Figure 4.1 increases as the universe expands, and we can express the rate of this increase in terms of a :

$$\dot{V} = \frac{dV}{dt} = \frac{d}{dt} \left[\frac{4}{3}\pi R(t)^3 \right] = \frac{d}{dt} \left[\frac{4}{3}\pi a(t)^3 \right] = 4\pi a^2 \dot{a}$$

We can write the internal energy of the fluid as the product of its volume and energy density: $E = V(t)\rho(t)c^2$. Remember that we have specified $R(t)/a(t) = 1$, which means we can write:

$$\begin{aligned} \dot{E} &= \frac{dE}{dt} = \frac{d}{dt} [V(t)\rho(t)c^2] = \frac{d}{dt} \left[\frac{4}{3}\pi a^3 \rho(t) c^2 \right] \\ &= \frac{4}{3}\pi c^2 (3a^2 \dot{a} \rho + a^3 \dot{\rho}) \end{aligned} \quad (4.7)$$

Substituting these expressions for \dot{V} and \dot{E} in Equation 4.6, we find that

$$dE + PdV = \frac{4}{3}\pi c^2 (3a^2 \dot{a} \rho + a^3 \dot{\rho}) + 4\pi a^2 \dot{a} P = 0$$

All that remains is to rearrange and simplify this expression: we arrive at the fluid equation as you will see it written in many cosmology textbooks.

The fluid equation

$$\dot{\rho} + 3\frac{\dot{a}}{a} \left(\rho + \frac{P}{c^2} \right) = 0 \quad (4.8)$$

The fluid equation shows us that $\dot{\rho}$ is governed by separate terms that depend on the current values for the density ρ and the pressure P of the fluid. The density term, which is proportional to $\dot{a}\rho/a$, just describes a dilution effect. If no new fluid is created as the universe expands, then conservation of energy means that the density must decrease. The interpretation of the pressure term is slightly more subtle. It arises because the fluid has lost internal energy by doing work as the volume it occupies expands.

- Explain why the pressure term in the fluid equation does *not* imply non-conservation of energy as the universe expands.
- Energy is still conserved, because although the *internal* energy of the fluid has decreased, this is exactly balanced by a corresponding increase in its gravitational potential energy, because the universe's expansion increases the distance between separate elements of the fluid.

This increase in gravitational potential might be easier to intuit by conceiving of the fluid as a finite number of massive particles distributed evenly throughout space. Isotropic expansion increases the distance between the particles, which increases the gravitational potential energy of the system.

4.1.3 The acceleration equation

In this section we will derive a third equation called the **acceleration equation** that explicitly describes the change in the rate at which a universe expands. The acceleration equation is not independent of the Friedmann and fluid equations but, as we will discover later in this chapter, having an equation that describes an accelerating scale factor is particularly useful for modelling our own Universe.

To derive the acceleration equation we will perform a series of operations to combine the Friedmann and fluid equations. We start by taking the time derivative of the Friedmann equation:

$$\begin{aligned} \frac{d}{dt} \left[\left(\frac{\dot{a}}{a} \right)^2 \right] &= 2 \frac{\dot{a}}{a} \frac{a\ddot{a} - \dot{a}^2}{a^2} = \frac{d}{dt} \left(\frac{8\pi G}{3} \rho - \frac{kc^2}{a^2} \right) \\ &= \frac{8\pi G}{3} \dot{\rho} + 2 \frac{\dot{a}}{a} \frac{kc^2}{a^2} \end{aligned} \quad (4.9)$$

Next, we replace $\dot{\rho}$ in Equation 4.9 using the fluid equation. After simplifying and rearranging the result, we obtain:

$$\frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a} \right)^2 = -4\pi G \left(\rho + \frac{P}{c^2} \right) + \frac{kc^2}{a^2}$$

The Friedmann equation can be used to replace $(\dot{a}/a)^2$ and a final reorganisation yields the acceleration equation.

The acceleration equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + \frac{3P}{c^2} \right) \quad (4.10)$$

Equation 4.10 tells us that increasing the density of the fluid in our model universe increases its gravitational influence and reduces the rate at which the scale factor a increases. A slightly less intuitive implication is that increasing the pressure of the fluid also *slows* the universe's expansion.

We are used to conceiving of positive pressure as a mechanism for transferring forces and inflating objects like balloons; however, forces are only transmitted when a pressure *gradient* is present. The balloon inflates because its internal pressure is higher than the ambient atmospheric pressure so there is a pressure gradient across its boundary. In a homogeneous universe there can be no pressure gradients and isotropic expansion preserves homogeneity. In the formalism of Einstein's field equations, the pressure is found to contribute to the overall **stress energy** of the Universe and, ultimately, it is stress energy that curves spacetime to produce gravitation.

4.2 Modelling the contents of the Universe

The two Friedmann equations and the fluid equation that you learned about in Section 4.1 are fundamental to the study of cosmology. Together, they allow us to model the histories and predict the futures of any universe that is described by the Robertson–Walker metric.

In this section you will learn how cosmologists use the Friedmann and fluid equations to model the evolution of the Universe. We will investigate how the scale factor behaves in homogeneous, isotropically expanding universes and how the contents of those universes affect the way they evolve.

4.2.1 The equation of state

The Friedmann equation and the fluid equation describe the co-evolution of a and ρ in a homogeneous, isotropically expanding universe. However, these equations can only be solved once we specify how the pressure term in the fluid equation evolves with time.

An equation that relates the pressure of a fluid to other quantities, such as its density or temperature, is called the fluid's **equation of state**. It is likely that you have met equations of state before. For example, the ideal gas law, which relates the pressure of a gas to its temperature and the volume it occupies, is also an equation of state.

The standard approach in cosmology is to model the contents of the Universe as a **perfect fluid**. Perfect fluids have particularly simple equations of state that depend *only* on their density and pressure.

Equation of state for a perfect fluid

$$P(\rho) = w\rho c^2 \quad (4.11)$$

The new parameter, w , in Equation 4.11 is the **equation of state parameter**. The evolution of universes containing different types of perfect fluid can be described by choosing different values of w to specify $P(\rho)$ before solving the Friedmann and fluid equations.

In subsequent sections we will examine the behaviour of model universes for which the curvature parameter $k = 0$. As noted in Section 4.1.1, such models are often called flat universes (you may also see them described as ‘flat cosmologies’). This is not because the spacetime within them is completely flat everywhere: mass and energy still produce local spacetime curvature in flat universes. However, on very large scales, the spatial parts of their metrics are Euclidean.

We will also focus on universes that contain the same types of fluid that we know exist in the real Universe. You will discover that the densities of different fluids evolve at different rates as the model universes expand, which has interesting implications for the history of our own Universe.

4.2.2 Models for matter

Cosmologists use the term ‘matter’ to refer to any fluid comprising non-relativistic material that has negligible or zero pressure. For such fluids, the equation of state parameter $w_m = 0$, where the subscript ‘m’ indicates that this property pertains to a matter-like fluid (or ‘matter fluid’). The equation of state has a very simple form in this scenario.

Equation of state for matter

$$P_m = 0 \quad (4.12)$$

You may be wondering whether there are any materials in the real Universe that behave like this idealised description of matter? The answer is yes! In fact, we know from astrophysical observations that around 50% of the baryonic matter in the Universe currently exists as a rarefied plasma of ions called the warm-hot intergalactic medium (or WHIM for short), which is distributed in the space between galaxies.

The WHIM contains some of the hottest gas in the Universe: its typical temperature is $T_{\text{WHIM}} \approx 10^7 \text{ K}$. Exercise 4.2 asks you to show that even this superheated plasma is effectively pressureless and contains only non-relativistic particles.

Exercise 4.2

The baryons in the WHIM are so diffuse that they almost never interact with one another, which means that they can be very accurately modelled as an ideal gas.

The ideal gas law can be written as

$$P = \frac{k_B T}{\mu} \rho \quad (4.13)$$

where P is the pressure of the gas, T is the temperature, μ is the mean mass of the particles in the gas and k_B is the Boltzmann constant. The typical kinetic energy of non-relativistic particles with temperature T is simply $3k_B T/2$.

Use this information to answer the following questions.

- (a) A fluid can be considered to behave like matter if its equation of state parameter $w \ll 1$. Show that $w < 10^{-6}$ for a plasma containing only protons that have a temperature $T_p \approx T_{\text{WHIM}}$, therefore implying that the WHIM behaves like our cosmological model for a matter fluid.

- (b) Assuming that the protons with $T_p \approx T_{\text{WHIM}}$ are non-relativistic, determine their typical squared velocity $\langle v_p^2 \rangle$.
- (c) Based on the result of part (b), was the assumption that the protons are non-relativistic appropriate?

A large fraction of the baryons in the Universe have now coalesced to form dense structures with non-negligible internal pressures. However, modelling the matter content of the Universe as having $w_m = 0$ is still a good approximation. Furthermore, in Chapter 1 you learned that baryonic matter contributes less than one-fifth of the total matter content of the Universe. The remaining $\sim 80\%$ is non-baryonic dark matter, which interacts so weakly with itself and the rest of the Universe that it can be very accurately described as being a pressureless fluid. This means that dark matter is also consistent with the description of matter that began this section.

Using Equation 4.12, and setting $P = 0$ in Equation 4.8, the fluid equation for matter can be written as

$$\frac{\dot{\rho}_m}{\rho_m} = -3\frac{\dot{a}}{a} \quad (4.14)$$

In the following example, you will see how this form of the fluid equation can be solved to predict how the density of matter evolves as the Universe expands, and how it depends on the value of the scale factor.

Example 4.2

Solve Equation 4.14 to show that $\rho_m \propto a^{-3}$.

Solution

Begin by multiplying both sides of Equation 4.14 by an infinitesimal time interval dt to obtain

$$\frac{1}{\rho_m} d\rho_m = -3\frac{1}{a} da$$

Integrating both sides yields

$$\begin{aligned} \int \frac{1}{\rho_m} d\rho_m &= -3 \int \frac{1}{a} da \\ \ln \rho_m &= -3 \ln(a) + C \end{aligned}$$

Now just apply the exponential function to both sides and simplify to find the required proportionality:

$$\begin{aligned} \exp(\ln \rho_m) &= \exp[-3 \ln(a) + C] \\ \implies \rho_m &= \exp[-3 \ln(a)] \exp(C) \\ &= a^{-3} \exp(C) \\ &\propto a^{-3} \end{aligned}$$

To fix the constant of proportionality, we need to specify some boundary conditions. We can use the convention that we described in Chapter 3 (Exercise 3.4) that sets $a = 1$ at the present time, when the age of the Universe is $t = t_0$. If we define $\rho_{m,0} = \rho_m(t_0)$ to be the present density of matter in the Universe, then we can relate the scale factor and the density of matter as follows.

Scale factor dependence of matter density

$$\rho_m = \frac{\rho_{m,0}}{a^3} \quad (4.15)$$

Now that we know how matter density depends on the scale factor, we can solve the Friedmann equation to determine how a evolves with time. Substituting for ρ_m in Equation 4.3 and assuming a flat universe with $k = 0$, we find:

$$\dot{a}^2 = \frac{8\pi G}{3a} \rho_{m,0} \quad (4.16)$$

It is important to remember that this form of the Friedmann equation only applies in universes that are both matter-dominated and spatially flat on large scales. In contrast, Equation 4.15 is true for the matter component of any universe. In the following exercise you will use Equation 4.16 to predict how the scale factor evolves with time in a universe that *is* flat and matter-dominated.

Exercise 4.3

Solve Equation 4.16 to show that in a flat, matter-dominated universe

$$a(t) = \left(\frac{t}{t_0} \right)^{2/3} \quad (4.17)$$

where $t_0 = (6\pi G \rho_{m,0})^{-1/2}$ is the time at the present day for a flat, matter-dominated universe.

We can use Equations 4.15 and 4.17 to predict how the density ρ_m evolves in a flat, matter-dominated universe.

Time dependence of matter density

$$\rho_m(t) = \frac{\rho_{m,0}}{a^3} = \rho_{m,0} \frac{t_0^2}{t^2} \quad (4.18)$$

4.2.3 Models for radiation

When cosmologists talk about radiation-like fluids (or ‘radiation fluids’) in the context of model universes, they mean fluids consisting of relativistic particles, with $v \approx c$. Examples of real particles that are always relativistic include photons and neutrinos.

Photons are massless, and are therefore relativistic by definition: in a vacuum, their speed exactly equals c . Although most particle physicists now believe that neutrinos must have a non-zero mass,[†] this mass is so small that all neutrinos move at speeds very near c .

For relativistic radiation fluids, whose properties are indicated in this chapter with a subscript ‘r’, the equation of state parameter $w_r = 1/3$, which is reflected in the resulting expression for their equation of state.

Equation of state for radiation

$$P_r = \frac{1}{3}\rho_r c^2 \quad (4.19)$$

As we did for the matter-dominated model in Section 4.2.2, we can substitute for P in Equation 4.8 to obtain the fluid equation for radiation:

$$\frac{\dot{\rho}_r}{\rho_r} = -4\frac{\dot{a}}{a}$$

Now, following exactly the same steps that we took for the universe that was matter-dominated, and introducing $\rho_{r,0} = \rho_r(t = t_0)$ as the current density of radiation in this radiation-only universe, we find the following dependence.

Scale factor dependence of radiation density

$$\rho_r = \frac{\rho_{r,0}}{a^4} \quad (4.20)$$

As before, we can write a simplified Friedmann equation that assumes $k = 0$ in this universe:

$$\dot{a}^2 = \frac{8\pi G}{3a^2}\rho_{r,0} \quad (4.21)$$

We know from our experience of solving Equation 4.16 that the solutions to equations like Equation 4.21 take the form of power laws with $a(t) \propto t^q$. To determine the value of the index q for a universe filled with radiation

[†]The belief that neutrinos have mass is related to the observation of neutrinos changing flavour or ‘oscillating’ as they propagate from their sources to the point at which they are detected. According to the Standard Model of particle physics, such oscillations would not be possible if neutrinos were truly massless.

we substitute this trial power law solution into Equation 4.21 and rearrange to find that

$$qt^{q-1} = \sqrt{\frac{8\pi G\rho_{r,0}}{3}}t^{-q} \quad (4.22)$$

For the equality in Equation 4.22 to hold, the powers of t must be equal on both sides of the equation, which means that $q = 1/2$, and our full solution for the radiation-only model becomes

$$a(t) = \left(\frac{t}{t_0}\right)^{1/2}$$

Exercise 4.4

Derive an expression for the present age t_0 of a flat, radiation-only universe in terms of $\rho_{r,0}$. Your expression should be analogous to the one for t_0 in Exercise 4.3.

Substituting for a in Equation 4.20, we can now predict how the radiation density evolves with time.

Time dependence of radiation density

$$\rho_r(t) = \frac{\rho_{r,0}}{a^4} = \rho_{r,0} \frac{t_0^2}{t^2} \quad (4.23)$$

Let's explore why the density of radiation (i.e. Equation 4.20) decreases faster than the density of matter (Equation 4.15) as the Universe expands.

Figure 4.2 shows two spheres with respective volumes V_m and V_r and constant co-moving radii. In other words, V_m and V_r are expanding along with the Universe. V_m contains a fixed number of massive particles and V_r contains a fixed number of photons. As the Universe expands, the volumes of the spheres increase by a factor a^3 and the *number* density of the particles and photons they contain decreases by the same factor.

The individual energies of the massive particles are not affected by the expansion, so their overall energy density $\rho_m c^2 \propto a^{-3}$, as we saw in Example 4.2. The situation is different for the photons in V_r . The energy of an individual photon is proportional to its wavelength, λ , such that $E = hc/\lambda$, where h is the Planck constant. As the Universe expands, the wavelengths of the photons in V_r get stretched by a factor a , which decreases their energy by the same factor. So each individual photon has a times less energy after the Universe has expanded and, when combined with the decrease in photon number density, we find that the energy density is reduced by a factor of a^4 , as shown in Equation 4.23.

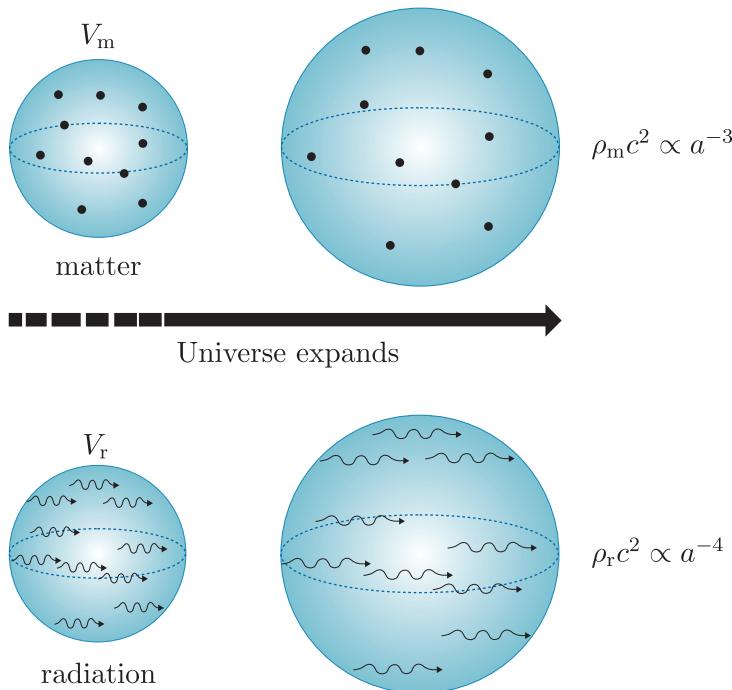


Figure 4.2 As the Universe expands the number densities of the photons in V_r and the massive particles in V_m both decrease by a factor $\propto a^{-3}$, but the photons lose energy corresponding to an additional factor of a as their wavelengths are stretched by the expansion.

4.2.4 Mixture models

The single-fluid universes that we investigated in the previous two sections are not very realistic. We know that the real Universe contains both matter and radiation, so in this section we will investigate the properties of more realistic ‘mixed’ universes that contain both relativistic and non-relativistic components.

To keep things simple in our discussions, we will assume that the two types of fluid in our mixed-model universes do not interact. However, in the *real* Universe, there are several ways that matter and radiation can interact and exchange energy. Two examples are given below.

- Matter is continuously being converted into radiation by nuclear fusion reactions in the cores of stars.
- In the space close to black hole event horizons, massive particles like protons and electrons can be accelerated to ultra-relativistic speeds. When the kinetic energy of these particles starts to dominate their rest mass energy, they start to behave like a radiation fluid.

However, stellar processes and extreme particle acceleration only affect fluids in very *localised* regions. Radiation will eventually be reabsorbed and converted into the kinetic energy of massive particles, and accelerated particles will eventually lose their kinetic energy and become non-relativistic again.

On the very large scales that all the models discussed in this chapter consider, relativistic and non-relativistic fluids can be considered as being separate. One *possible* mechanism for interaction between matter and radiation fluids on larger scales is that of matter–antimatter annihilation. When a massive particle encounters its antiparticle, the two are annihilated and their rest mass energy is converted entirely to radiation. In reality, the matter content of the Universe today is dominated by ‘normal’ matter, and antimatter is relatively scarce. This makes interactions between matter and antimatter rare enough that this mechanism for converting matter to radiation can normally be neglected on cosmological scales.

Our assumption that matter and radiation fluids do not interact means that separate fluid equations can be used to describe the evolution of their densities independently. The solutions to these independent fluid equations are identical to those we found for the single-component universes, and we would still find that $\rho_m = \rho_{m,0}/a^3$ and $\rho_r = \rho_{r,0}/a^4$.

Solving the Friedmann equation for multi-component universes is not as straightforward. Both fluids exert a gravitational influence, so the evolution of the scale factor depends on both of their densities at the same time. This means that we cannot construct independent Friedmann and fluid equations for matter and radiation in this scenario. Instead, we must use the total density, $\rho_{\text{total}} = \rho_m + \rho_r$, to write:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho_{\text{total}} - \frac{kc^2}{a^2} = \frac{8\pi G}{3}(\rho_m + \rho_r) - \frac{kc^2}{a^2} \quad (4.24)$$

We will not attempt to solve Equation 4.24 to find $a(t)$ and $\rho_{m,r}(t)$ for arbitrary values of ρ_m and ρ_r . Even if we assume that $k = 0$, the solution is quite complicated when the values of ρ_m and ρ_r are comparable.

However, it is useful to consider solutions when the density of one component dominates over that of the other. In fact, we know that there have been times in the history of the Universe when its density has been dominated by radiation, and other times when it has been dominated by matter.

How would a spatially flat universe containing matter and radiation evolve in the radiation-dominated case? We can assume that a evolves as if there were no matter present, so $a(t) \propto t^{1/2}$, and the density of its radiation-fluid component ρ_r^r is

$$\rho_r^r(t) \propto a^{-4} \propto t^{-2}$$

For the matter component, its density ρ_m^r still evolves such that $\rho_m^r(a) \propto a^{-3}$, but because the evolution of a is dominated by ρ_r^r , the *time* evolution of ρ_m^r is different from that in a matter-only universe:

$$\rho_m^r(t) \propto a^{-3} \propto t^{-3/2}$$

So in a radiation-dominated universe, ρ_m^r decreases more slowly than ρ_r^r . This means that a flat, two-component universe that is initially radiation-dominated, with $\rho_r^r \gg \rho_m^r$, must eventually evolve to become

The superscript ‘r’ indicates that this mixed universe is radiation-dominated, while the subscript ‘r’ implies we are describing a property of the radiation-fluid component.

matter-dominated. Conversely, a universe that also contains only matter and radiation but is currently matter-dominated must once have been dominated by radiation, regardless of the present-day densities $\rho_{m,0}$ and $\rho_{r,0}$.

When a two-component universe becomes matter-dominated, we can neglect the influence of the radiation component. Following the same argument as we did for the radiation-dominated case, we find that

$$\begin{aligned} a(t) &\propto t^{2/3} \\ \rho_m^m &\propto a^{-3} \propto t^{-2} \\ \rho_r^m &\propto a^{-4} \propto t^{-8/3} \end{aligned}$$

So in the matter-dominated case, ρ_r^m decreases faster than ρ_m^m . This means that once this model universe becomes matter-dominated, it will stay that way forever.

Until the late 1990s, most cosmologists believed that a two-component model involving just matter and radiation provided a good description for the contents of the Universe. In Section 4.4 we will explain why this is no longer the case, and explore the implications of what is one of the most poorly understood properties of our Universe.

Similar to the previous explanation, the superscript ‘m’ is used to indicate properties of fluids in a matter-dominated universe, while the subscript ‘m’ implies we are describing a property of the matter-fluid component.

4.3 Density and curvature

In this section we will rewrite the Friedmann equation in a way that explicitly reflects the link between the spatial curvature k of the Universe and the densities of its various components.

4.3.1 The critical density

In Section 4.1.1 we demonstrated that the value of k is spatially and temporally invariant. Once we know k , the co-evolution of a and ρ is completely specified by the Friedmann equation, the fluid equation, and the equations of state for each of the Universe’s component fluids. Conversely, if we *know* the value of a and how fast the Universe is expanding, we can ask what its density must be in order for k to have a particular value.

Cosmologists often make use of a quantity called the **critical density** (ρ_c), which is the density required for the Universe to be spatially flat if the Hubble parameter, $H(t) = \dot{a}/a$, has a particular value. In a flat Universe $k = 0$ and $\rho = \rho_c$ by definition, so the Friedmann equation becomes

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho_c(t) \quad (4.25)$$

Making the substitution for $H(t)$, we can rearrange Equation 4.25 to obtain an expression for the critical density.

The critical density

$$\rho_c(t) = \frac{3H(t)^2}{8\pi G} \quad (4.26)$$

Note that $\rho_c(t)$ is a time-dependent quantity that varies in proportion to the Hubble parameter squared as the Universe expands or contracts. By replacing $H(t)$ with the Hubble constant H_0 , we can also write down the present critical density of the Universe:

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} \quad (4.27)$$

The following example should help to give you a physical intuition for the magnitude of the critical density today.

Example 4.3

Answer the following questions assuming that $H_0 = 68 \text{ km s}^{-1} \text{ Mpc}^{-1}$, and otherwise using values provided in the table of constants.

- (a) What is the value of the present-day critical density, $\rho_{c,0}$, in kg m^{-3} ?
- (b) How many protons per m^3 does this value correspond to?
- (c) If the mass of a typical galaxy is 10^{11} M_\odot (where M_\odot is the mass of the Sun), how many galaxies per cubic megaparsec are required to make the Universe flat today?

Solution

- (a) To evaluate $\rho_{c,0}$ using Equation 4.27, we need to convert the value we are given for the Hubble constant into SI units. One megaparsec equals $3.086 \times 10^{22} \text{ m}$, so:

$$H_0 = 68 \text{ km s}^{-1} \text{ Mpc}^{-1} = \frac{68 \times 1000 \text{ m}}{3.086 \times 10^{22} \text{ m}} \text{ s}^{-1} = 2.2 \times 10^{-18} \text{ s}^{-1}$$

Now we can evaluate Equation 4.27:

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} = \frac{3 \times (2.2 \times 10^{-18})^2}{8\pi \times 6.674 \times 10^{-11}} \text{ kg m}^{-3} = 8.7 \times 10^{-27} \text{ kg m}^{-3}$$

- (b) The mass of a proton is $1.673 \times 10^{-27} \text{ kg}$, so the present critical density equates to approximately 5 protons per cubic metre.
- (c) To answer the final part of the question we need to convert $\rho_{c,0}$ to solar masses per cubic megaparsec. We use the fact that $1 \text{ M}_\odot = 1.99 \times 10^{30} \text{ kg}$ to deduce:

$$\begin{aligned} \rho_{c,0} &= 8.7 \times 10^{-27} \text{ kg m}^{-3} \\ &= 8.7 \times 10^{-27} \times \frac{(3.086 \times 10^{22})^3}{1.99 \times 10^{30}} \text{ M}_\odot \text{ Mpc}^{-3} \\ &= 1.3 \times 10^{11} \text{ M}_\odot \text{ Mpc}^{-3} \end{aligned}$$

So the present critical density corresponds to approximately one galaxy per cubic megaparsec. This value turns out to be quite a bit larger than the observed number density of galaxies in the Universe.

4.3.2 Density parameters

It is important to keep in mind that the critical density is not necessarily the true density of a universe. If a universe has *non-zero* curvature then its true density ρ will be different from its critical density ρ_c .

To quantify this difference, cosmologists define a dimensionless ratio called the **density parameter**.

Density parameter

$$\Omega(t) = \frac{\rho}{\rho_c} \quad (4.28)$$

- In general, $\Omega(t)$ is time-dependent and varies as a universe evolves, but under what circumstances would a universe have a time-invariant value of Ω ?
- In flat universes, which are defined by having $k = 0$, Ω does not depend on time. Remember that k is time-invariant, so if $k = 0$ then ρ always equals ρ_c , and Ω is *always* exactly 1.

Using the definition of ρ_c in Equation 4.26, we can rewrite the Friedmann equation (Equation 4.3) in terms of the density parameter:

$$H(t)^2 = \frac{8\pi G}{3}\rho_c\Omega - \frac{kc^2}{a^2} = H(t)^2\Omega - \frac{kc^2}{a^2}$$

Rearranging this expression to isolate the density parameter, we obtain the Friedmann equation in terms of Ω :

$$\Omega - 1 = \frac{kc^2}{H(t)^2 a^2} \quad (4.29)$$

For universes with non-zero k , the value of $\Omega(t)$ varies as the universe evolves, but we can infer from Equation 4.29 that its *sign* never changes. If $k > 0$ then the density of the universe must always be larger than the critical density and, vice versa, if $k < 0$ then the density of the universe must always be less than ρ_c . Using the convention that the present-day scale factor $a(t_0) = 1$, Equation 4.29 can be expressed in terms of the current critical density and the Hubble constant as

$$\Omega_0 - 1 = \frac{\rho_0}{\rho_{c,0}} - 1 = \frac{kc^2}{H_0^2} \quad (4.30)$$

The following example shows how Equation 4.30 can be used to constrain the curvature of the real Universe at earlier times in its history.

Example 4.4

Current observational evidence indicates that $|\Omega_0 - 1| < 0.003$. Assuming that the Universe contains a mixture of matter and radiation, show that the value of Ω must have been even closer to 1 in the past.

Solution

We can rewrite Equation 4.29 as

$$\Omega - 1 = \frac{kc^2}{\dot{a}^2}$$

The value of k never changes, so to prove that $\Omega - 1$ was *smaller* in the past, we must show that \dot{a} was *larger* in the past.

We showed in Sections 4.2.2 and 4.2.3 that $a \propto t^{2/3}$ for matter-dominated universes and $a \propto t^{1/2}$ for universes dominated by radiation. In both cases, \dot{a} decreases with time and must have been larger in the past. These results from these hypothetical scenarios imply that if the real Universe contains only matter and radiation, then its rate of expansion should also be slowing down and would have been larger in the past.

Ω is commonly expressed as a sum of density parameters for each component of a universe's contents. For a universe containing matter and radiation, we would write:

$$\Omega = \frac{\rho}{\rho_c} = \frac{\rho_m}{\rho_c} + \frac{\rho_r}{\rho_c} = \Omega_m + \Omega_r \quad (4.31)$$

To simplify the appearance of equations and calculations, cosmologists also define an effective density parameter for curvature:

$$\Omega_k = -\frac{kc^2}{a^2 H^2} \quad (4.32)$$

Ω_k contains no more *physical* information than k does, but using this convention we can rewrite the Friedmann equation very compactly as

$$\Omega + \Omega_k = 1 \quad (4.33)$$

4.4 The cosmological constant

4.4.1 Introducing Λ

When Einstein published his general theory of relativity in 1915, astronomical observations had only revealed a fraction of the Universe that is visible to modern astronomers. Based on the observational data that were available at the time, Einstein quite justifiably believed that the Universe was static, with $\dot{a} = 0$. It would be almost 15 years before astronomers established that the Universe was expanding by measuring the recession velocities of distant galaxies.

Einstein realised that if the Friedmann equation was correct (Equation 4.3 or, equivalently, Equation 4.4), and the Universe has non-zero energy density, then \dot{a} could only be exactly zero at one instant in its history. Even if the Universe *was* momentarily static, the gravitational influence of its contents would make it start to collapse. We saw in Section 4.1.1 that the only possible solutions to Equation 4.3 when a universe is *not* empty describe a universe that expands forever, or one that expands initially and then recollapses.

To reconcile this apparent disagreement between his theory and the observations that were available in 1915, Einstein chose to modify his field equations. His modifications introduced a new repulsive term in the Friedmann equation that maintained $\dot{a} = 0$ by exactly balancing the gravitational attraction of the matter and radiation in the Universe. Einstein called this new term the **cosmological constant** and it is normally written in equations as Λ (the Greek upper-case letter lambda).

As the name suggests, the value of Λ is identical at all points in space and does not evolve with time. With this new term, the Friedmann equation can be rewritten as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \quad (4.34)$$

The following exercise is designed to give you some practice using the Friedmann equation with a cosmological constant term.

Exercise 4.5

Einstein introduced a positive cosmological constant to enable the existence of a static, *non-empty* universe. Show that such a universe must have a positive curvature by considering whether each term in Equation 4.34 is positive or negative in this context.

Once astronomers had established that the Universe was indeed expanding, Einstein happily abandoned the cosmological constant. He had considered its somewhat arbitrary introduction to be an affront to the mathematical beauty of his theory and famously referred to this as his ‘greatest blunder’.

In 1930, Arthur Eddington published a paper showing that even with a cosmological constant, static universes are inherently unstable. Perturbing the matter or radiation density by a tiny amount would upset the precise balance between Λ and the gravitating components of such a universe. Once that happened, the universe would begin to expand or contract according to whether the initial density perturbation was negative or positive.

The resultant change in the scale factor would amplify the initial density perturbation and accelerate the expansion or contraction. Remember that Λ is constant by definition, and so the more the density perturbation

grows, the less Λ is able to balance its gravitational influence. The inevitable result of any small density perturbation in Einstein's static universe is runaway expansion or collapse.

Despite its abandonment during the early twentieth century, the cosmological constant has enjoyed a renaissance in recent years. In the late 1990s, astronomers measured the brightness of distant supernovae and found that they appeared fainter than expected. In Chapter 5, you will learn that this has been interpreted as evidence that the expansion of the Universe is actually accelerating. Moreover, the observed acceleration seems to be exactly consistent with the expected effect of a real cosmological constant.

If the cosmological constant is real, then what generates it? At the time of writing, the physical origin of Λ and its true nature remain to be firmly established. Many cosmologists use the name 'dark energy' to refer to the unknown phenomenon that is responsible for the cosmological constant. You will learn more about dark energy and some popular theories for its physical origin later in this module.

4.4.2 Implications of the cosmological constant

A variety of observations strongly suggest that a cosmological constant term contributes to the current energy density of the real Universe. In this section we will explore how the introduction of Λ affects some of the equations that we derived earlier in this chapter.

First, to derive the acceleration equation for a universe with a cosmological constant, we can follow the steps in Section 4.1.3. Replacing Equation 4.3 with Equation 4.34 in that derivation gives

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda c^2}{3} \quad (4.35)$$

We can now see explicitly that positive values of Λ increase the value of \ddot{a} . In other words, a positive cosmological constant acts like a repulsive force that increases the rate at which the Universe expands.

It is worth noting that there is no mathematical reason for Λ to be positive. Given our current lack of knowledge about the true origins of Λ , there is no theoretically motivated physical reason either! However, the effect of Λ on the behaviour of \ddot{a} in the acceleration equation, coupled with the fact that the Universe appears to be accelerating in its expansion, strongly suggests that $\Lambda > 0$ in the real Universe.

A fluid description of the cosmological constant

Cosmologists often choose to model the cosmological constant as a third fluid component of the Universe, with its own *constant* energy density ($\epsilon_\Lambda = \rho_\Lambda c^2$) and pressure (P_Λ). If we adopt this approach, we can rewrite Equation 4.34 as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}(\rho + \rho_\Lambda) - \frac{kc^2}{a^2} \quad (4.36)$$

Comparing this expression with Equation 4.34, we see that

$$\frac{8\pi G}{3}\rho_\Lambda = \frac{\Lambda c^2}{3}$$

and so

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G} \quad (4.37)$$

We can also write down a separate fluid equation (see Equation 4.8) for Λ :

$$\dot{\rho}_\Lambda + 3\frac{\dot{a}}{a}\left(\rho_\Lambda + \frac{P_\Lambda}{c^2}\right) = 0 \quad (4.38)$$

For the cosmological constant, the equation of state parameter $w_\Lambda = -1$, which implies that it has a negative effective pressure. Remember that fluids with positive pressures exert an attractive gravitational effect. The fact that $P_\Lambda < 0$ is yet another way to see that Λ generates a repulsive force that counteracts gravity.

Using Equation 4.37 we can also express the density parameter, which you met in Section 4.3.2, for Λ :

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} = \frac{\Lambda c^2}{8\pi G} \frac{8\pi G}{3H(t)^2} = \frac{\Lambda c^2}{3H(t)^2} \quad (4.39)$$

Using this expression together with Equation 4.32, the Friedmann equation can be rewritten as

$$\Omega = \Omega_m + \Omega_r + \Omega_\Lambda = 1 - \Omega_k \quad (4.40)$$

4.4.3 Models of the Universe with a cosmological constant

The introduction of a cosmological constant term adds a new degree of freedom to the Friedmann equation and permits many more possible histories and futures of the Universe. In this section we will explore some of these models and see how they differ from those for universes with $\Lambda = 0$.

To help our exploration, it will be useful to rewrite the Friedmann equation in terms of the density parameters that were introduced in Section 4.3.2. The following example illustrates how to derive the expression we need.

Example 4.5

Show that the Friedmann equation can be written as

$$\frac{H^2}{H_0^2} = \frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \Omega_{\Lambda,0} - \frac{(\Omega_0 - 1)}{a^2} \quad (4.41)$$

Solution

Starting from the familiar Newtonian form of the Friedmann equation (Equation 4.3), we can use Equation 4.30 to replace k :

$$H^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} = \frac{8\pi G}{3}\rho - \frac{H_0^2(\Omega_0 - 1)}{a^2}$$

Dividing this expression by H_0^2 and then using Equation 4.27 we find:

$$\frac{H^2}{H_0^2} = \frac{8\pi G}{3H_0^2}\rho - \frac{\Omega_0 - 1}{a^2} = \frac{\rho}{\rho_{c,0}} - \frac{\Omega_0 - 1}{a^2}$$

Next, we express ρ as the sum of the densities of the Universe's components:

$$\frac{H^2}{H_0^2} = \frac{\rho_m + \rho_r + \rho_\Lambda}{\rho_{c,0}} - \frac{\Omega_0 - 1}{a^2}$$

By definition, $\rho_\Lambda = \rho_{\Lambda,0}$ is constant, and in Section 4.2 we showed that $\rho_m = \rho_{m,0}a^{-3}$ and $\rho_r = \rho_{r,0}a^{-4}$, so:

$$\begin{aligned} \frac{H^2}{H_0^2} &= \frac{\rho_{m,0}a^{-3} + \rho_{r,0}a^{-4} + \rho_{\Lambda,0}}{\rho_{c,0}} - \frac{\Omega_0 - 1}{a^2} \\ &= \frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \Omega_{\Lambda,0} - \frac{\Omega_0 - 1}{a^2} \end{aligned}$$

The first three terms on the right-hand side of Equation 4.41 separately describe how the energy density of each component influences the evolution of a , while the final term encapsulates the influence of curvature. This form of the Friedmann equation is very helpful when exploring the current behaviour of model universes that include multiple components, and in particular those with a cosmological constant.

Figure 4.3 illustrates the possible behaviours of a simple model in which radiation contributes a very small proportion of the total energy density. If $\Omega_{r,0} \approx 0$, then Equation 4.41 becomes

$$\begin{aligned} \frac{H^2}{H_0^2} &= \frac{\Omega_{m,0}}{a^3} + \Omega_{\Lambda,0} - \frac{\Omega_0 - 1}{a^2} \\ &= \frac{\Omega_{m,0}}{a^3} + \Omega_{\Lambda,0} + \frac{1 - \Omega_{m,0} - \Omega_{\Lambda,0}}{a^2} \end{aligned} \quad (4.42)$$

This expression tells us that the behaviours of our simplified model depend on the values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ at different points in the $\Omega_{m,0}$ – $\Omega_{\Lambda,0}$ plane. The curves in Figure 4.3 divide this plane based on different characteristics that a universe described by Equation 4.42 can exhibit. The remainder of this section considers these different characteristics in more detail.

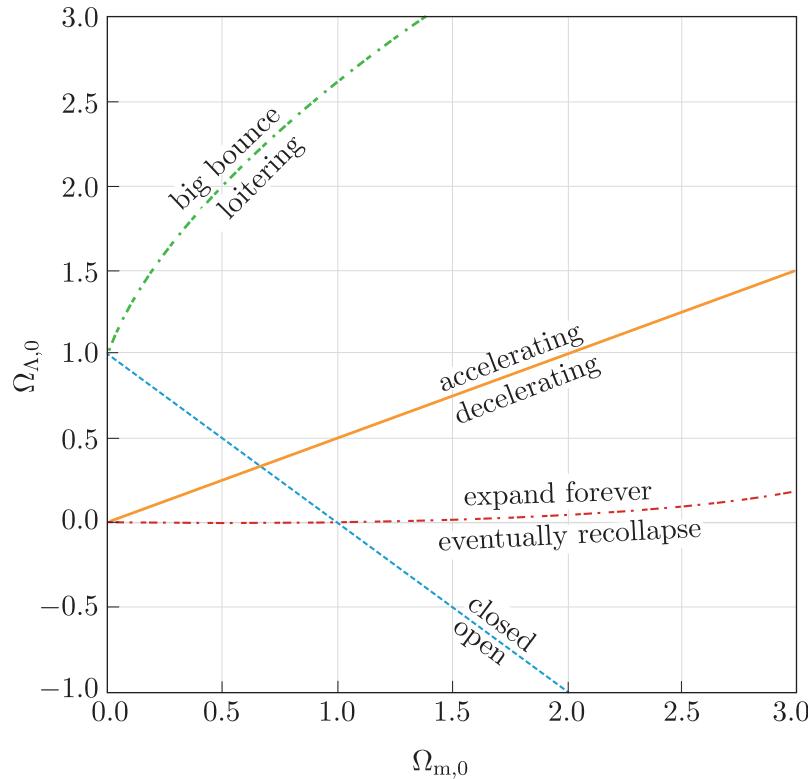


Figure 4.3 Possible characteristics of a model universe that contains matter and a cosmological constant, but in which the density of radiation is negligible.

Open, closed and flat universes

The easiest question to answer about a model universe is whether it is flat, open or closed. This depends solely on the sign of $\Omega_k = 1 - \Omega_{m,0} - \Omega_{\Lambda,0}$. If $\Omega_{m,0} > 1 - \Omega_{\Lambda,0}$ then $\Omega_k < 0$, so $k > 0$ and the universe is closed; such universes lie above the blue dashed line in Figure 4.3. Conversely, if $\Omega_{m,0} < 1 - \Omega_{\Lambda,0}$ then the universe is open and would lie below the blue dashed line in the figure. Flat universes lie exactly on the blue dashed line, having $\Omega_{m,0} = 1 - \Omega_{\Lambda,0}$.

- Is a universe with positive Λ and total density $\rho > \rho_c$ guaranteed to end by collapsing to zero size in a big crunch? (*Hint:* you may need to refer to Figure 4.3 to determine your answer.)
- No. If $\rho > \rho_c$ then $k > 0$, $\Omega_k < 0$, and the universe must be closed, and lie above the blue dashed line on Figure 4.3. The plot shows us that there are a large number of universe types with positive Λ that are closed and that also lie above the red dot-dashed line. These universes will not collapse: rather they will expand forever.

Without a cosmological constant, the curvature k , the total density ρ and the ultimate fate of the Universe are all intimately related. If $\Lambda = 0$, then any universe that has $\rho > \rho_c$ has positive k and must end in a big crunch, while any universe that has $\rho \leq \rho_c$ has zero or negative k and will expand forever. However, the *inclusion* of a cosmological constant in our model for the Universe destroys this simple relationship, and we can now describe closed universes that expand forever as well as open universes that eventually recollapse.

Accelerating, decelerating and coasting universes

The next question we can ask is whether the expansion of the model universe we are investigating is currently accelerating or decelerating. To answer this question, we will derive a quantity called the **deceleration parameter**.

The deceleration parameter

$$q_0 = -\frac{\ddot{a}(t_0)}{a(t_0)} \frac{1}{H_0^2} = -\frac{a(t_0)\ddot{a}(t_0)}{\dot{a}(t_0)^2} \quad (4.43)$$

From Equation 4.43 we see that the value of q_0 is negative if \ddot{a} is positive, and therefore the rate of expansion of the universe (\dot{a}) is currently increasing. Conversely, q_0 is positive if \ddot{a} is instead decreasing.

Our derivation of the deceleration parameter starts by writing down the first three terms of the Taylor expansion of the scale factor evaluated at the present time:

$$a(t) = a(t_0) + \dot{a}(t_0)(t - t_0) + \frac{1}{2}\ddot{a}(t_0)(t - t_0)^2 + \mathcal{O}(3) \quad (4.44)$$

The notation $\mathcal{O}(3)$ is a shorthand that is used to refer to any terms in the series that are proportional to powers of $(t - t_0)$ greater than 3. For values of t close to the present age of the Universe $(t - t_0) \ll 1$, and we expect that those terms represented by $\mathcal{O}(3)$ will be negligible. Next, we divide all terms by $a(t_0)$, which allows us to express the coefficients in terms of the Hubble constant, and simultaneously provides a way to define the deceleration parameter:

$$\frac{a(t)}{a(t_0)} = 1 + H_0(t - t_0) - \frac{q_0}{2}H_0^2(t - t_0)^2 + \mathcal{O}(3) \quad (4.45)$$

Comparing the coefficients of $(t - t_0)^2$ in Equations 4.44 and 4.45, we find:

$$\frac{1}{2} \frac{\ddot{a}(t_0)}{a(t_0)} = -\frac{q_0}{2} H_0^2$$

Rearranging this expression gives us the definition of the deceleration parameter q_0 in Equation 4.43.

Example 4.6

Show that if the energy density of a model universe is dominated by matter and a cosmological constant, then the deceleration parameter in that universe can be expressed as

$$q_0 = \frac{\Omega_{m,0}}{2} - \Omega_{\Lambda,0} \quad (4.46)$$

Solution

Start by writing Equation 4.35, assuming that $t = t_0$, $\rho = \rho_m$ and $P = P_m = 0$:

$$\frac{\ddot{a}(t_0)}{a(t_0)} = -\frac{4\pi G}{3} \left(\rho_{m,0} + \frac{3P_m}{c^2} \right) + \frac{\Lambda c^2}{3} \quad (4.47)$$

Dividing by H_0^2 and using Equation 4.27 we obtain

$$\begin{aligned} \frac{1}{H_0^2} \frac{\ddot{a}(t_0)}{a(t_0)} &= -\frac{4\pi G}{3H_0^2} \rho_{m,0} + \frac{\Lambda c^2}{3H_0^2} \\ &= -\frac{1}{2} \frac{\rho_{m,0}}{\rho_{c,0}} + \frac{\Lambda c^2}{3H_0^2} \end{aligned}$$

To derive Equation 4.46 we need to recognise, using Equation 4.43, that the left-hand side of the previous expression is equal to $-q_0$. Then, using the definitions of Ω from Equation 4.28 (evaluated for the matter component at the present time) and Ω_Λ from Equation 4.39, we can write:

$$q_0 = -\frac{1}{H_0^2} \frac{\ddot{a}(t_0)}{a(t_0)} = \frac{\Omega_{m,0}}{2} - \Omega_{\Lambda,0}$$

From Equation 4.46 we can straightforwardly infer that when $\Omega_{\Lambda,0} > \Omega_{m,0}/2$ then $q_0 < 0$, and so \ddot{a} is positive. Universes that fulfil this criterion lie above the solid orange line in Figure 4.3. Conversely, if $\Omega_{\Lambda,0} < \Omega_{m,0}/2$ for a universe then it would lie below the solid orange line in the figure, and its rate of expansion would be slowing down. Universes whose expansion is neither accelerating nor decelerating lie exactly on the solid orange line, and are often described as ‘coasting’ universes.

‘Loitering’ universes

By carefully fine-tuning the values of $\Omega_{\Lambda,0}$ and $\Omega_{m,0}$ it is possible for expanding universes with negative curvature to enter a state in which a hardly evolves at all. These solutions, where the density of matter is almost able to overcome a positive cosmological constant, are called ‘loitering’ universes. In Figure 4.3 they lie below, but very close to, the green dot-dashed line in the upper left-hand corner of the plot. An observer in such a universe might easily mistake it for a static universe unless they could make sufficiently precise measurements to detect the almost imperceptible expansion of spacetime.

Loitering universes eventually evolve away from their quasi-static state when their cosmological constant starts to dominate the gravitational effect of their matter content. Once that happens, a starts to grow exponentially and the universe will keep expanding forever.

‘Big bounce’ universes

Perhaps the strangest type of expanding universes that become possible when $\Lambda > 0$ are those in which there was no big bang, and in which a was never zero! In Figure 4.3 such universes lie above the green dot-dashed line. They all have positive curvature.

Let’s consider the case of one that is initially Λ -dominated and contracting, with $H = \dot{a}/a < 0$. This model universe is positively curved, so Equation 4.40 tells us that the third term in Equation 4.42, which can be expressed as Ω_k/a^2 , must be negative. If $\Omega_{\Lambda,0}$ and $\Omega_{m,0}$ are both positive and have appropriate values, it is possible for this third term in Equation 4.42 to become sufficiently dominant in time to halt the contraction (reducing H to zero) and allow an expansion phase with $a > 0$ to start. Universes like this are sometimes referred to as ‘big bounce’ universes, in reference to the rebounding evolution of their scale factors.

4.5 Summary of Chapter 4

- The **Friedmann equation** describes how the rate of the Universe’s expansion evolves with time, and how that evolution depends on the Universe’s contents:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \quad (\text{Eqn 4.34})$$

- The **fluid equation** describes how the densities of the different components of the Universe evolve as the Universe itself expands or contracts:

$$\dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{P}{c^2}\right) = 0 \quad (\text{Eqn 4.8})$$

- The **acceleration equation** can be derived from the Friedmann and fluid equations, and can be used to determine whether the expansion of the Universe is speeding up or slowing down:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda c^2}{3} \quad (\text{Eqn 4.35})$$

- The contents of the universe can be modelled as non-interacting **perfect fluids** with different **equations of state**.
- In matter-dominated universes the energy density decreases over time as the universe expands, and the scale factor increases such that

$$\rho_m(t) = \frac{\rho_{m,0}}{a^3} = \rho_{m,0} \frac{t_0^2}{t^2} \quad (\text{Eqn 4.18})$$

- In radiation-dominated universes the energy density decreases over time as the universe expands and the scale factor increases such that

$$\rho_r(t) = \frac{\rho_{r,0}}{a^4} = \rho_{r,0} \frac{t_0^2}{t^2} \quad (\text{Eqn 4.23})$$

- The geometric curvature of the Universe is time-invariant. If the curvature is known then the Friedmann equation, the fluid equation and the equations of state for all components of a model universe can be used to predict how that universe's density and expansion rate will evolve throughout its lifetime.
- The **critical density** ρ_c refers to the total density a universe must have to be spatially flat, with its curvature parameter k exactly equal to zero. In general, ρ_c varies as the universe expands or contracts.
- For each component of the Universe a corresponding **density parameter** can be defined. This time-varying quantity is equal to the ratio of the component's density to the critical density of the Universe. It is also possible to define an effective density parameter corresponding to the effect of the spatial curvature, k .
- There is observational evidence that the real Universe contains a **cosmological constant** component, Λ , that counteracts the gravitational influence of the other components and can be modelled as a perfect fluid with a constant, negative pressure.
- The cosmological constant greatly increases the number of possible evolutionary histories that a model universe can follow. The exact path followed depends in a detailed way on the relative densities of the universe's different components, including the cosmological constant, and the universe's geometry.

Chapter 5 Measuring cosmological parameters

In the previous chapter you learned how to describe the evolution of the scale factor $a(t)$ for different model universes, which can have different contents and different intrinsic properties, like curvature. In this chapter you will learn how cosmologists can determine the properties of the *real* Universe by measuring the different parameters that we find in the Friedmann equations.

Our strategy will involve carefully measuring the distances to different celestial objects, which will let us determine how $a(t)$ has evolved over time. Once we know $a(t)$, we can use the Friedmann equations to constrain the energy densities of matter, radiation and the cosmological constant (Λ), as well as the overall curvature of the Universe.

Objectives

Working through this chapter will enable you to:

- list and describe four ways to define the distance to objects in an expanding universe
- discuss how the apparent distances to objects depend on the parameters of the Friedmann equation
- explain how cosmologists use several types of celestial objects and phenomena to measure distances in the Universe
- summarise the different techniques that cosmologists apply to infer distances to these objects
- explain how cosmologists use observations of distant objects to measure the Hubble constant, and thereby determine the *current* expansion rate of the Universe
- explain how cosmologists use observations of distant objects to measure density parameters in the Friedmann equation, which determine the expansion *history* of the Universe.

5.1 Defining distance

In this chapter you will learn how measurements of the distances to celestial objects allow cosmologists to determine the values of cosmological parameters like the Hubble parameter, H_0 , and the density parameters $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$. However, in an expanding universe, there are several ways that the distances to remote objects like distant galaxies can be defined, and which definition to choose often depends on the scientific question that you are trying to answer.

You have already learned about co-moving coordinates, which measure distances that remain unchanged as the Universe expands. In this section you will also learn about the proper distance – which measures the length of spatial geodesics – as well as the luminosity distance and the angular diameter distance, which are both related to the proper distance but based on observable properties of the objects whose distance they quantify.

5.1.1 Proper distance

The formal definition of the **proper distance** between two objects is the length of a spatial geodesic connecting their positions at a *particular* cosmic time, t , in the history of the Universe, when its scale factor was $a(t)$. An observer located at the origin at cosmic time t can compute the proper distance to a distant galaxy (or any distant object) by integrating the component of the Robertson–Walker metric in Equation 3.10 that depends directly on the radial coordinate r . If distance to the galaxy is measured in *radial coordinates* as R , then the corresponding proper distance d_p is defined by Equation 5.1.

Proper distance

$$d_p(t) = a(t) \int_0^R \frac{dr}{\sqrt{1 - kr^2}} \quad (5.1)$$

The proper distance is closely related to the **co-moving distance**, d_c . Specifically, at cosmic time t the two are related via the scale factor, such that:

$$d_p(t) = a(t)d_c \quad (5.2)$$

Note that d_c itself does *not* depend on cosmic time. This is because it is defined in terms of the co-moving coordinate grid that you read about in Section 3.3.1. In that earlier discussion, the scale of the co-moving coordinate grid – in other words, the numerical coordinate values associated with each of the grid lines – was left unspecified. In *principle* we could choose to label the lines in the evenly spaced coordinate grid with any set of evenly spaced numerical values.

However, in *practice*, cosmologists conventionally fix the co-moving coordinate scale by defining the present value of the scale factor to be $a(t_0) = 1$, which implies that:

$$d_p(t_0) = a(t_0)d_c = d_c \quad (5.3)$$

Equation 5.3 reveals that an important consequence of choosing $a(t_0) = 1$ is that the *present* proper distance between two objects equals the co-moving distance d_c between them, and vice-versa.

The proper distance turns out to be a very useful quantity for theoretically understanding the Universe, but unfortunately it cannot be measured in

any practical way. Cosmologists must rely on observations of light from distant objects to infer how far away they are. Bearing this in mind, let's find an expression for the present proper distance in terms of quantities that are more closely related to observable properties of distant objects that *can* actually be measured.

We start by considering a beam of light propagating from a distant galaxy, at radial coordinate R , to a detector on the Earth. The photons in the beam travel along null geodesics in the radial direction (i.e. $ds^2 = 0$), and so the Robertson–Walker metric can be written as:

$$ds^2 = c^2 dt^2 - a(t)^2 \frac{dr^2}{1 - kr^2} = 0 \quad (5.4)$$

We can collect together the terms that depend on t and those that depend on r as follows:

$$\frac{c^2 dt^2}{a(t)^2} = \frac{dr^2}{1 - kr^2} \quad (5.5)$$

If we assume that the photons from the galaxy were emitted at time t_{em} and detected some later at time t_{obs} , and that they travelled from $r = R$ to $r = 0$, then we can write that:

$$c \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)} = \int_0^R \frac{dr}{\sqrt{1 - kr^2}} \quad (5.6)$$

The right-hand side of this equation is just the definition of d_c , or $d_p(t)/a(t)$ (compare with Equation 5.1). In the situation where $t_{\text{obs}} = t_0$, and therefore $a(t) = a(t_0) = 1$, we arrive at an expression for the proper distance that relates directly to measurements of cosmic time:

$$d_p(t_0) = c \int_{t_{\text{em}}}^{t_{\text{obs}}} \frac{dt}{a(t)} \quad (5.7)$$

In the past, when $a(t) < 1$, then $d_p < d_c$. In the future the opposite will be true (i.e. $d_p > d_c$) for as long as the Universe continues to expand. Later in this chapter, you will learn that Equation 5.7 is more straightforwardly linked to quantities that can actually be observed than the original definition of proper distance (Equation 5.1), and is therefore practically useful. The following example shows how to calculate an important quantity called the **horizon distance**, which is equal to the maximum proper distance that photons could have travelled since the big bang.

Example 5.1

The horizon distance d_{hor} defines the proper radius of a sphere (i.e. the proper distance from its centre to its outer surface), which is also called the **particle horizon**. This sphere surrounds an observer at a particular cosmic time and encompasses their entire observable universe.

Find an expression for d_{hor} in a universe that is spatially flat and is matter-dominated.

Solution

To find d_{hor} we set $t_{\text{em}} = 0$ in Equation 5.7, which lets us compute the proper distance that photons have been travelling since the Universe began. We cannot observe objects at greater distances than this, because light from them has not yet had time to reach us. In other words, they are outside of our observable universe.

In Section 4.2.2 we showed that $a(t) = (t/t_0)^{2/3}$ in a matter-dominated universe, and $t = 0$ at the big bang, so Equation 5.7 becomes:

$$d_{\text{hor}}(t_0) = ct_0^{2/3} \int_0^{t_0} t^{-2/3} dt \quad (5.8)$$

Now, we simply evaluate the integral to find:

$$\begin{aligned} d_{\text{hor}}(t_0) &= ct_0^{2/3} \left[3t^{1/3} \right]_0^{t_0} \\ &= 3ct_0^{2/3} t_0^{1/3} \\ &= 3ct_0 \end{aligned}$$

This result shows that the present-day horizon distance in a flat, matter-dominated universe is three times further than the distance light could have travelled since the big bang!

Example 5.1 demonstrates the remarkable result that the Friedman equations can model universes in which the horizon distance exceeds ct_0 . The example assumed a simple matter-only model, but if we performed a similar calculation for the constituents of the real Universe, we would still find that $d_{\text{hor}}(t_0) > ct_0$.

- Does this assertion about the real Universe mean that photons from the particle horizon have travelled with superluminal velocity (i.e. faster than the speed of light) to reach us? If not, how are the findings of Example 5.1 possible?
- No, these photons have not travelled faster than c . Remember that $d_{\text{hor}}(t_0)$ specifies the distance to the particle horizon *today*. Light that is now reaching us from the horizon completed most of its journey in the past, when the Universe – and therefore the horizon distance – was much smaller.

Equation 5.7 in its current form is still not directly useful. If we knew t_{em} and the functional form for $a(t)$, we could use Equation 5.7 to compute the proper distance that photons arriving at Earth today have travelled since the time when they were emitted. However, even if we assume a model for $a(t)$, we do not yet have any way of inferring a value for the time t_{em} .

Fortunately, we can rewrite Equation 5.7 in terms of a quantity that we *can* actually measure – the redshift z of the photons arriving from the distant galaxy. The following exercise asks you to derive an expression that explicitly links d_p and z .

Exercise 5.1

By using the Friedmann equation in the form shown in Equation 4.41 and adopting the standard convention that $a(t_0) = 1$, show that

$$d_p(t_0) = \frac{c}{H_0} \int_0^{z_{\text{em}}} \frac{dz}{E(z)} \quad (5.9)$$

is equivalent to Equation 5.7. In Equation 5.9, z_{em} is the redshift of the photons we observe at time t_0 from a source at a proper distance $d_p(t_0)$, and

$$E(z) = \sqrt{\Omega_{m,0}(1+z)^3 + \Omega_{r,0}(1+z)^4 + \Omega_{\Lambda,0} + \Omega_{k,0}(1+z)^2}$$

where $\Omega_{k,0}$ is the present-day value of Ω_k . (*Hint:* you may need to refer back to an equation in Chapter 3 that relates a and z in the context of emitted and observed light.)

If we can measure the redshift of photons from a distant object then Equation 5.9 becomes very useful, because it directly relates $d_p(t_0)$ to the values of H_0 , $\Omega_{m,0}$, $\Omega_{r,0}$, $\Omega_{\Lambda,0}$ and Ω_k . In Sections 5.1.2 and 5.1.3 you will learn about two methods that cosmologists use to estimate $d_p(t_0)$ for distant objects, based on their observable properties. By using such estimates, obtained for a large number of distant objects, we can use Equation 5.9 to constrain the parameters in the Friedmann equation.

Different sources of redshift

At low redshifts, Equation 5.9 reduces to the Hubble–Lemaître law that you learned about in Chapter 1. Now we see that the distance that appears in Equation 1.2 is in fact the proper distance evaluated at the present time, t_0 :

$$z = \frac{H_0 d_p(t_0)}{c} \quad (5.10)$$

It is important to remember that the redshift that appears in Equation 5.10 is the *cosmological redshift*. This is the redshift that we would measure for a galaxy that was static with respect to the Hubble flow, with fixed co-moving coordinates. This distinction has implications for how accurately we can actually measure H_0 , because objects in the Universe are not static. Every galaxy has its own **peculiar velocity**, which adds a Doppler-shift component to the redshift that we measure.

There is no way to determine what proportion of a galaxy's *measured redshift* results from its peculiar velocity. If we tried to estimate H_0 using observations of a *single* galaxy, then our answer would be shifted by an unknown amount from the true value. However, if we apply the cosmological principle that the Universe is homogeneous and isotropic, then we can expect the *average* peculiar velocity of *many* galaxies to be very close to zero. Therefore, if we can accurately estimate the proper

distances to a large number of galaxies with known redshifts, then we can compute an average value of H_0 that should be very close to the true value.

Measuring galaxies' redshifts using spectroscopy is relatively straightforward with modern astronomical instruments. However, collecting sufficiently precise distance measurements for a large sample of distant galaxies is a significant technical challenge. Later in this chapter you will learn about the sophisticated techniques that cosmologists use to measure distances on cosmological scales.

5.1.2 Luminosity distance

Suppose we somehow *know* that the luminosity of a distant object is L . Objects with known luminosity are often referred to as **standard candles**, and you will see some examples later in this chapter. If we observe one of these standard candles and measure its flux to be F , then we can calculate an estimate for its distance from us using a quantity called the **luminosity distance**.

Luminosity distance

$$d_L = \sqrt{\frac{L}{4\pi F}} \quad (5.11)$$

It is important to recognise that the luminosity distance for an object at a particular redshift is not necessarily the same as its (present-day) proper distance, $d_p(t_0)$. Consider the scenario shown in Figure 5.1.

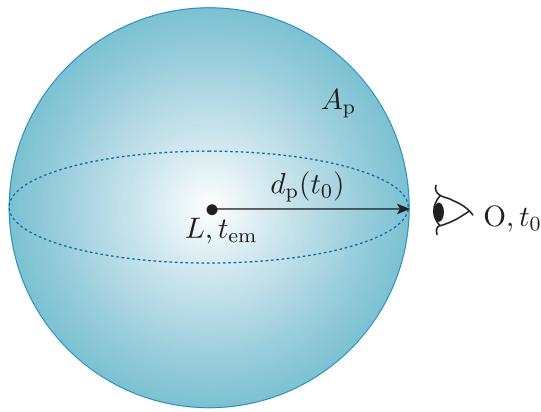


Figure 5.1 At time t_0 an observer located at point O , at a proper distance $d_p(t_0)$, measures the light from a standard candle of luminosity L . Photons that were emitted earlier, at time t_{em} , are distributed over a sphere with proper surface area $A_p = 4\pi d_p(t_0)^2$.

A present-day observer at point O studies a standard candle with luminosity L , at a co-moving distance of $d_p(t_0)$. They measure at time t_0

the photons that were emitted isotropically at time t_{em} . If the current proper distance between the observer and the light source is $d_p(t_0)$, then photons that were emitted at t_{em} will be distributed over the surface of a sphere with radius $d_p(t_0)$ and proper surface area $A_p = 4\pi d_p(t_0)^2$. For a static universe this means that the observed flux at O will be $F = L/A_p$.

- Does the value of F depend on the value of the curvature parameter k ? Briefly explain your answer.
- Yes it does, because k appears in the definition of the proper distance d_p (Equation 5.1), which itself appears in the expression for A_p .

In a static, spatially flat universe, $d_p(t_0)$ is equal to the luminosity distance. However, d_L and $d_p(t_0)$ are *not* equal in the real Universe, where both cosmic expansion and curvature can modify the flux that we observe.

For simplicity we will start by considering the impact of cosmic expansion in a spatially flat universe. In such a universe, two related effects act to reduce the observed flux of distant luminous objects relative to what would be expected in a static universe.

To understand these two effects, it is important to recognise that both flux and luminosity are defined as *rates* of energy transfer: they describe the amount of energy that is emitted or received within a particular time interval. In Chapter 3 you saw how cosmic expansion produces a cosmological redshift that reduces the frequency, and therefore the *energy*, of photons by a factor of $1+z$ as they propagate through the Universe.

A similar effect decreases the *rate* at which photons arrive compared to the static universe case. The following example explores this effect.

Recall that the energy E and frequency ν of a photon are related via $E = h\nu$, where h is the Planck constant.

Example 5.2

Consider two photons that follow identical spatial trajectories and arrive at O in Figure 5.1 when the scale factor is $a(t_0)$.

Assuming that the photons were *emitted* at time t_{em} , when the scale factor was $a(t_{\text{em}})$, show that the interval between the photons' arrivals, δt_0 , is longer than the interval between their emissions, δt_{em} , by a factor $(1+z)$, where z is the redshift of the standard candle.

Solution

To show that this is the case, we can use a very similar argument to the steps used to derive the relation between a and z in Section 3.3.3.

Let the arrival times of the photons be t_0 and $t_0 + \delta t_0$, and assume that their corresponding emission times are t_{em} and $t_{\text{em}} + \delta t_{\text{em}}$. We can use Equation 5.7 to relate the two photons' emission and reception times to the corresponding scale factors:

$$c \int_{t_{\text{em}}}^{t_0} \frac{dt}{a(t)} = c \int_{t_{\text{em}} + \delta t_{\text{em}}}^{t_0 + \delta t_0} \frac{dt}{a(t)}$$

To separate terms that depend on t_0 and t_{em} we can subtract a third integral from both sides, following steps (i) to (iv) below.

$$\begin{aligned}
 \text{(i)} \quad & c \int_{t_{\text{em}}}^{t_0} \frac{dt}{a(t)} - c \int_{t_{\text{em}}+\delta t_{\text{em}}}^{t_0} \frac{dt}{a(t)} = c \int_{t_{\text{em}}+\delta t_{\text{em}}}^{t_0+\delta t_0} \frac{dt}{a(t)} - c \int_{t_{\text{em}}+\delta t_{\text{em}}}^{t_0} \frac{dt}{a(t)} \\
 \text{(ii)} \quad & c \int_{t_{\text{em}}}^{t_{\text{em}}+\delta t_{\text{em}}} \frac{dt}{a(t)} = c \int_{t_0}^{t_0+\delta t_0} \frac{dt}{a(t)} \\
 \text{(iii)} \quad & \frac{1}{a(t_{\text{em}})} \int_{t_{\text{em}}}^{t_{\text{em}}+\delta t_{\text{em}}} dt = \frac{1}{a(t_0)} \int_{t_0}^{t_0+\delta t_0} dt \\
 \text{(iv)} \quad & \frac{\delta t_{\text{em}}}{a(t_{\text{em}})} = \frac{\delta t_0}{a(t_0)}
 \end{aligned}$$

We can now use Equation S1 to show:

$$\frac{\delta t_0}{\delta t_{\text{em}}} = 1 + z$$

Therefore we have shown that the interval between photon arrival from an object at redshift z will be a factor of $1 + z$ larger in an expanding universe than in a static one.

Accounting for both effects of the Universe's expansion on the observed flux, the relationship between the flux and luminosity of an object at redshift z in a universe where $k = 0$ is:

$$F = \frac{L}{4\pi(1+z)^2 d_p(t_0)^2}$$

By comparison with Equation 5.11, we find that an object's luminosity distance, its cosmological redshift and its current proper distance from Earth are related such that

$$d_L = (1+z) d_p(t_0) \tag{5.12}$$

In curved universes this relationship becomes more mathematically complex, but with a similar increasing divergence between the two distance measures at large redshifts.

Figure 5.2 compares curves of d_L and $d_p(t_0)$ versus redshift for three model universes with very different contents. Panel (a) shows d_L for a large range of redshifts between 0 and 10. In panel (b) d_L is plotted for much smaller redshifts, corresponding to much closer light sources. For comparison, both panels also show grey curves that represent the present day proper distance $d_p(t_0)$ as a function of redshift. The grey and coloured curves with the same line styles correspond to the same model universe assumptions. Note that d_L is always larger than the corresponding proper distance.

In all three cases, the luminosity distance can be used to accurately estimate the proper distance for nearby objects with $z \lesssim 0.04$. However, for more distant objects d_L consistently overestimates $d_p(t_0)$. In other words, objects in the high-redshift universe appear to be further away than they really are if we only use their apparent brightness to estimate their distance from us.

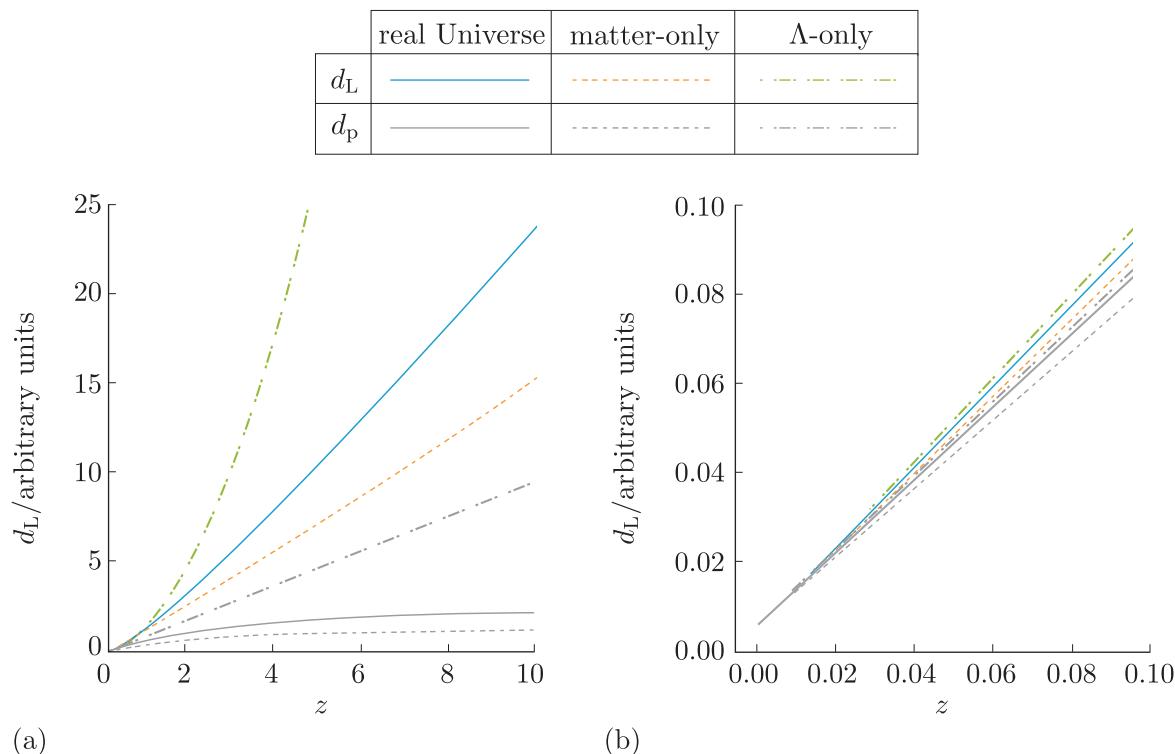


Figure 5.2 d_L (coloured lines) and d_p (grey lines) versus redshift for three model universes. Panel (a) shows a large range of redshifts out to $z = 10$, whereas panel (b) focuses on the nearby Universe.

Later in this chapter you will learn how measurements of the luminosity distance to standard candles can be used to constrain the cosmological parameters. However, this is not the only way that d_L can be useful. Astrophysicists and astronomers often want to know the intrinsic luminosities of distant celestial objects that are *not* standard candles. If they can measure an object's redshift z and its flux F , then Equations 5.9, 5.11 and 5.12 can be used to determine L .

- What extra assumptions must we make in order to calculate L once F and z have been measured?
- We need to assume a particular set of cosmological parameters so that we can evaluate $d_p(t_0)$ using Equation 5.9.

The luminosity distances to standard candles can be used to measure the cosmological parameters, but once those parameters have been measured, the luminosity distance allows the luminosities of *any* celestial light source to be inferred using its measured flux.

5.1.3 Angular diameter distance

Objects or phenomena in the Universe that are known to have a fixed physical size or length scale are called **standard rods**. You may also see the terms ‘standard ruler’ and ‘standard yardstick’ used to describe such objects.

Cosmologists define a quantity called the **angular diameter distance**, d_A , that relates the length l of a standard rod to its apparent angular size θ when we observe it. For the distant objects that we are focused on in this chapter, $\theta \ll 1$, so we can use the small angle approximation and define d_A as follows:

Angular diameter distance

$$d_A = \frac{l}{\sin \theta} \approx \frac{l}{\theta} \quad \text{for } \theta \text{ expressed in radians.} \quad (5.13)$$

The angular diameter distance would be equal to the current proper distance in a static, spatially flat universe. In the *real* Universe d_A and $d_p(t_0)$ are *not* equal, but they are related to each other such that

$$d_A = \frac{d_p(t_0)}{1+z} \quad (5.14)$$

To derive this relationship, we will consider the scenario shown in Figure 5.3. We will again assume a flat universe for simplicity.

The two ends of a standard rod are defined in terms of metric coordinates by the points $A = (t_{\text{em}}, r, \theta_A, \phi)$ and $B = (t_{\text{em}}, r, \theta_B, \phi)$. Photons emitted from A and B at time t_{em} propagate along null geodesics to reach an observer at the point O at time t_0 . The observed angle between the photons is θ_{obs} .

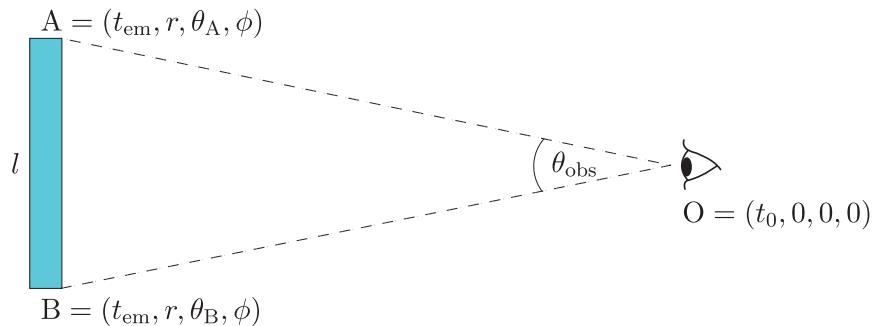


Figure 5.3 An observer O at the origin observes the light from points A and B at either end of a standard rod of length l . The observed angle between A and B is θ_{obs} .

Now, let's consider another observer who was located at O at time t_{em} , when the photons were *emitted* and the proper distance between O and the rod was $d_p(t_{\text{em}})$. Let θ_{AB} represent the angular size of the rod as seen by our second observer. If we assume that the length of the rod is constant and that $d_p(t_{\text{em}})$ is large enough that $\theta_{AB} \ll 1$, then we can write:

$$\theta_{AB} = \frac{l}{d_p(t_{\text{em}})} \quad (5.15)$$

The key insight is that because the expansion of the Universe is homogeneous and isotropic, the angle between the photons' trajectories must remain unchanged throughout their journey. Therefore, it must be true that $\theta_{AB} = \theta_{\text{obs}}$. Using this fact, we can rearrange Equation 5.15 to find an expression for l in terms of θ_{obs} , and use the relationships between $d_p(t)$, $d_p(t_0)$ and d_c from Equations 5.2 and 5.3 to find that:

$$\begin{aligned} l &= d_p(t_{\text{em}}) \theta_{AB} = d_p(t_{\text{em}}) \theta_{\text{obs}} \\ &= a(t_{\text{em}}) d_c \theta_{\text{obs}} = a(t_{\text{em}}) d_p(t_0) \theta_{\text{obs}} = \frac{1}{1+z} d_p(t_0) \theta_{\text{obs}} \end{aligned} \quad (5.16)$$

Finally, combining Equations 5.13 and 5.16 we recover the result stated in Equation 5.14:

$$d_A = \frac{l}{\theta_{\text{obs}}} = \frac{d_p(t_0)}{1+z}$$

As with Equation 5.12, this relation applies only to flat ($k = 0$) geometries, but a similar redshift dependence is present in more complex curved geometries.

Figure 5.4 compares curves of d_A and $d_p(t_0)$ versus redshift for three model universes with very different contents. The panels are analogous to those shown for the luminosity distance in Figure 5.2. Panel (a) shows d_A for a large range redshifts between 0 and 10; in panel (b), d_A is plotted for much smaller redshifts. Note that unlike the luminosity distance (which was always larger than $d_p(t_0)$) the angular diameter distance is always smaller than the corresponding proper distance.

For objects at low redshift ($z \lesssim 0.04$) the angular diameter distance provides a good approximation for the proper distance. For more distant objects, d_A consistently *underestimates* $d_p(t_0)$. To see why this is the case, remember that the co-moving distance r to the rod is constant so the integral in Equation 5.1 is also constant. Bearing this in mind, we can define z_{em} to be the observed redshift of photons that were emitted at time t_{em} and use Equation S1 to write

$$d_p(t_{\text{em}}) = \frac{a(t_{\text{em}})}{a(t_0)} d_p(t_0) = a(t_{\text{em}}) d_p(t_0) = \frac{d_p(t_0)}{1+z_{\text{em}}} = d_A \quad (5.17)$$

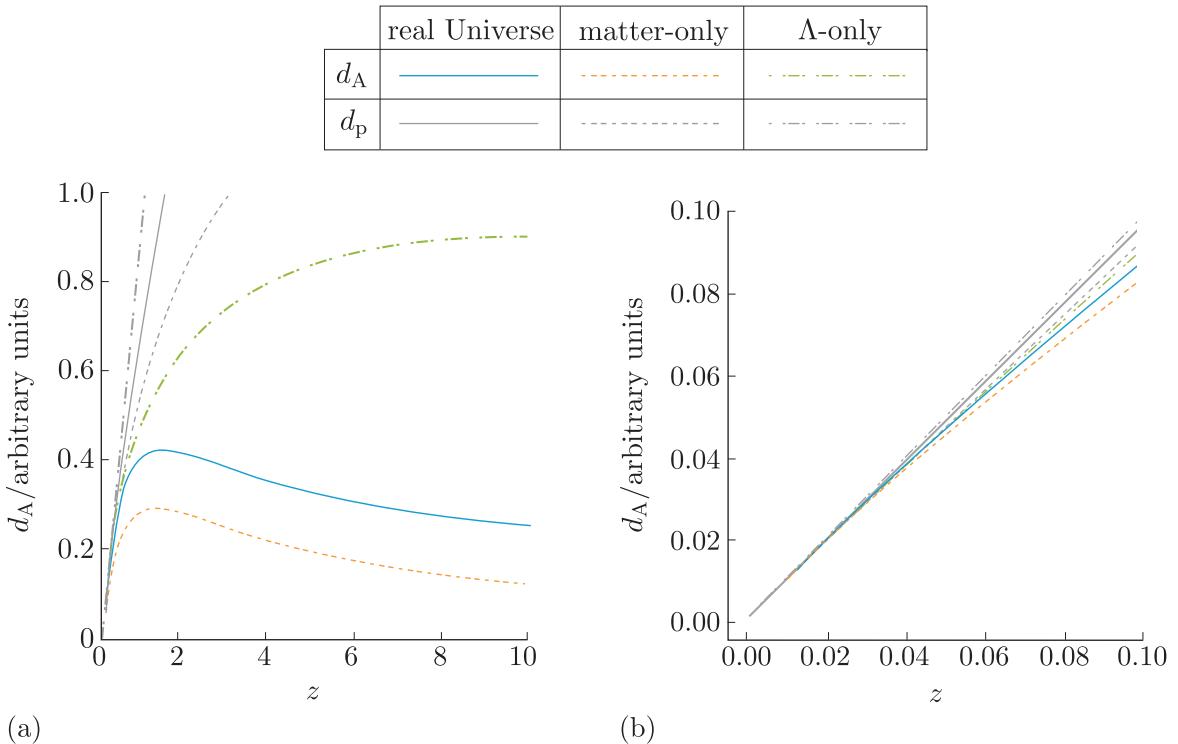


Figure 5.4 d_A (coloured lines) and d_p (grey lines) versus redshift across two ranges for three model universes. Panel (a) shows a large range of redshifts out to $z = 10$, whereas panel (b) focuses on the nearby Universe.

This shows that the present angular diameter distance is equal to the proper distance between the observer and the standard rod at the time t_{em} , when the observed photons were emitted and the scale factor was smaller.

Most remarkably, if a universe contains any component of matter or radiation, then the angular diameter distance only increases up to a certain critical redshift before objects start to appear *larger* as they move further away. If we use d_A as our distance estimate, then objects in the very distant Universe do not just appear closer than they really are: they also appear closer than other objects that actually *are* closer!

Exercise 5.2

Consider a population of objects that are simultaneously standard candles with known luminosity L and standard rods with known physical extent l . Show that the *observed* surface brightness Σ of these objects decreases rapidly with increasing redshift according to:

$$\Sigma \propto (1+z)^{-4} \quad (5.18)$$

(Hint: the surface brightness Σ is equivalent to the observed flux of the object, F_{obs} , divided by the solid angle it subtends on the sky.)

The result of Exercise 5.2 also has implications when observing celestial objects that are neither standard candles *nor* standard rods. The surface brightness of these objects is one of the key factors that determines whether or not they can be detected by a particular telescope, with a particular sensitivity. The fact that the surface brightness decreases so rapidly with redshift means that very powerful telescopes are required to detect and study high-redshift objects.

Just like luminosity distance, the angular diameter distance can be used to determine the properties of distant celestial objects.

- Which physical property of distant objects do you think d_A can be used to measure?
- It can be used to measure the physical *size* of distant objects.

If astronomers can measure the redshift and the *angular* size of an object, then Equations 5.9, 5.13 and 5.14 can be used to infer l . Again, this requires that we *assume* a particular set of cosmological parameters in order to calculate $d_p(t_0)$. Knowing the physical size of extended structures can help astrophysicists to understand the physical processes that produced them. For example, measuring the lengths of the jets launched by distant radio galaxies and the sizes of their radio lobes allows us to investigate the acceleration of charged particles close to the supermassive black holes that power distant active galactic nuclei.

5.2 Measuring distances

In the previous section you read about the different ways that cosmologists can define distances in the Universe. In this section we will introduce and explore some of the observational techniques that can be used to *measure* the distances to remote celestial objects. We will focus primarily on the ways that astronomers measure the distances to standard candles in the nearby and distant Universe. Then, in Section 5.3, you will learn how these distance measurements have allowed cosmologists to constrain the cosmological parameters.

5.2.1 Stellar parallax

It is possible to measure the distance to stars in the Milky Way and its satellite galaxies using very precise measurements of those stars' positions.

Consider a scenario in which an astronomer observes the apparent position of a nearby star relative to the positions of *much* more distant background objects, as illustrated schematically in Figure 5.5. Six months later, the Earth has completed half of its orbit around the Sun and its position has changed by a distance $2d_b$. The apparent position of the nearby star will also have shifted slightly when it is observed at this point, but the background objects will not seem to have moved. (In fact, the apparent positions of the distant objects will have shifted *very* slightly as well, but they are so far away that this shift is negligible.)

The shift in a nearby star's position relative to a fixed background is called **stellar parallax**, and can be used to determine the proper distance d_p to that star.

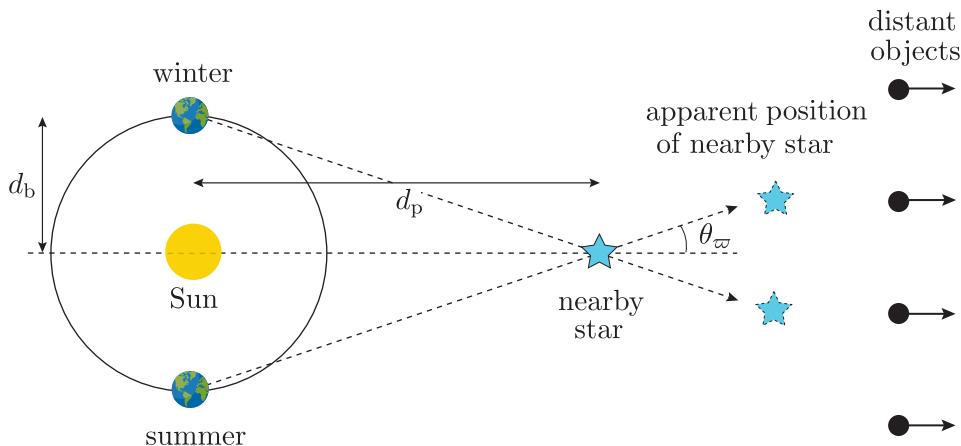


Figure 5.5 The apparent position of a nearby star (at distance d_p) is measured twice, six months apart, relative to much more distant background objects. During that interval the Earth's position has changed by $2d_b$. The small but measurable shift in the nearby star's apparent position is called its stellar parallax.

- If you were choosing a set of background sources to measure the stellar parallax of a star against, what properties would you want those sources to have?
- The most obvious property that the objects must have is that they are much further away than the target star. The fact that they must be so distant means that they should also be very luminous, so that we can actually detect them.

If we want to use the background sources to define a fixed frame of reference, then we need to be able to measure *their* positions very precisely without having to make any subjective decisions about them, for example where their centres are. For that reason, the ideal background sources are point-like, with no measurable angular extension. Distant quasars are celestial objects that fulfil all of these criteria and are frequently used when measuring distances using stellar parallax.

The angle between the two measured positions ($2\theta_\pi$) is related in the following way to the distance to the star (d_p) and the ‘baseline’ distance between the Earth’s location when the measurements were made ($2d_b$).

Calculating distances using stellar parallax

$$d_p = \frac{d_b}{\theta_\pi} \quad (5.19)$$

If the units of d_p and d_b are the same in Equation 5.19, then θ_ϖ has units of radians. However, lengths and angles in astronomy are often specified using a variety of different units. The following exercise will give you some practice using Equation 5.19 when a mixture of non-SI units are used.

Exercise 5.3

The Large Magellanic Cloud (LMC) is a satellite galaxy of the Milky Way that lies approximately $d_p = 163\,000$ light-years (ly) away. Assuming the nearby star illustrated in Figure 5.5 is somewhere in the LMC, estimate the value of θ_ϖ in arcseconds as seen from Earth.

The parsec unit

As you have seen in previous chapters, astronomers often state measured or estimated distances in units called parsecs, where $1\text{ pc} \approx 3.1 \times 10^{16}\text{ m}$. The parsec unit is defined, using Equation 5.19, as the value of d_p when d_b equals the radius of the Earth's orbit around the Sun and θ_ϖ is one arcsecond. The word parsec is actually a contraction of the term 'parallax second'.

The maximum distance that can be measured using stellar parallax is limited by the precision with which the positions of distant objects can be measured. Figure 5.6 shows an artist's impression of the *Gaia* satellite, which was launched in 2013 by the European Space Agency. *Gaia* has since been used to measure the positions and parallaxes of over 1 billion stars with an accuracy reaching ~ 24 microarcseconds!

Even with this astonishing precision, Equation 5.19 tells us that *Gaia* can only measure parallaxes for objects closer than $\sim 42\text{ kpc}$ (equivalent to $z \sim 10^{-5}$ via the Hubble-Lemaître law). Now look at the following example, which considers whether distances derived using stellar parallax – even using measurements made by *Gaia* – can be used to accurately determine the expansion rate of the Universe.

Example 5.3

- (a) The recession velocity of a star at time t is just the rate of change of proper radial distance:

$$v_r(t) = \frac{d}{dt}d_p(t)$$

Estimate the radial recession velocity caused by the expansion of the Universe for a star at the maximum distance that can be measured by *Gaia* using stellar parallax.

- (b) Compare this value with the estimated peculiar velocity of the Milky Way, which is $\sim 600\text{ km s}^{-1}$, and briefly comment on the implications for using nearby objects to estimate H_0 .

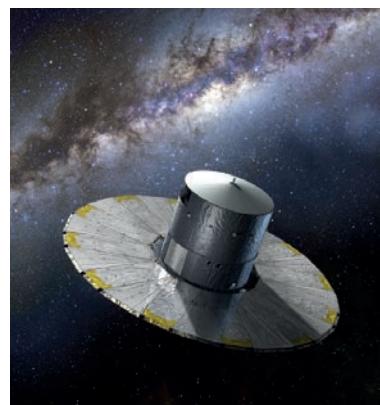


Figure 5.6 An artist's impression of the *Gaia* satellite.

Solution

- (a) We are only interested in the component of the recession velocity that is related to cosmic expansion, so we can write v_r in terms of the *co-moving* distance to the star, d_c , and relate this to the proper radial distance using the scale factor a and its time derivative:

$$v_r(t) = \dot{a}(t)d_c = \frac{\dot{a}(t)}{a(t)}d_p(t) = H(t)d_p(t) \quad (5.20)$$

$H(t)$ is the time-dependent Hubble parameter that appears in the Friedmann equation.

Now, we will consider a photon that is emitted by the star at time t_{em} and then detected by *Gaia* at time t_0 . We start by rearranging the relation between a and z (e.g. Equation S1) to isolate z :

$$z = \frac{a(t_0)}{a(t_{\text{em}})} - 1 \quad (5.21)$$

We have seen that *Gaia* can only measure parallaxes for very nearby objects, so we can assume $t_{\text{em}} \approx t_0$ and replace $a(t_{\text{em}})$ with its Taylor expansion around t_0 :

$$a(t_{\text{em}}) \approx a(t_0) + \dot{a}(t_0)(t_{\text{em}} - t_0) = a(t_0)[1 + H_0(t_{\text{em}} - t_0)]$$

Substituting for $a(t_{\text{em}})$ in Equation 5.21 we find:

$$z \approx \frac{a(t_0)}{a(t_0)[1 + H_0(t_{\text{em}} - t_0)]} - 1 = \frac{1}{1 - H_0(t_0 - t_{\text{em}})} - 1 \quad (5.22)$$

We can use a series expansion to further simplify our approximate expression for z .

$$\frac{1}{1 - x} \approx 1 + x + \mathcal{O}(x^2)$$

Here, we use the notation $\mathcal{O}(x^n)$ as shorthand to refer to terms that depend on powers of x greater than or equal to n . In this case, we expect that terms depending on powers of x greater than or equal to 2 will be negligible compared to terms that depend on smaller powers of x . It follows that:

$$z \approx \frac{1}{1 - H_0(t_0 - t_{\text{em}})} - 1 \approx H_0(t_0 - t_{\text{em}})$$

Our photon propagates at the speed of light, so if the current proper distance to the star is $d_p(t_0)$ then we can write:

$$c = \frac{d_p(t_0)}{t_0 - t_{\text{em}}}$$

So finally, our approximate expression for z becomes:

$$z \approx H_0(t_0 - t_{\text{em}}) \approx \frac{H_0 d_p(t_0)}{c}$$

We have therefore derived a form of the Hubble–Lemaître Law (Equation 1.2). Now we can use Equation 5.20 to relate z to the

current recession velocity due to cosmic expansion:

$$z \approx \frac{v_r(t_0)}{c}$$

This expression allows apparent superluminal recession velocities, but this does not violate special relativity. Remember that this recession velocity represents the expansion of spacetime, not the peculiar motions of objects within it, which *are* constrained to be less than c .

We know that the most distant objects for which *Gaia* can measure parallaxes would have cosmological redshifts as small as $z \sim 10^{-5}$.

Using this we compute a recession velocity:

$$v_r(t_0) \approx cz \approx 3 \text{ km s}^{-1} \quad (5.23)$$

- (b) This recession velocity value is just 0.5% of the Milky Way's peculiar velocity, which implies that any redshifts we measure for nearby objects are likely to be completely dominated by their peculiar motions. To measure velocities caused by the cosmic expansion, we need to measure distances to objects that are much further away.

5.2.2 Standard candles

Earlier in this chapter we described standard candles in the context of the luminosity distance. In this section you will learn about two types of celestial object that cosmologists use as standard candles, and which allow the distances to those objects and their host galaxies to be estimated.

Classical Cepheid stars

Classical Cepheids (CCs) are a group of periodic variable stars with luminosities that rise and fall on a regular timescale. They are named after a star called δ Cephei, which was the first example to be discovered. CCs are bright ($10^3 L_\odot \lesssim L \lesssim 10^4 L_\odot$) and massive ($4 M_\odot \lesssim M \lesssim 20 M_\odot$). They lie in a region of the Hertzsprung–Russell (H–R) diagram known as the **instability strip**, which is populated with several types of pulsating stars.

The variability of CCs is driven by instabilities in their outer atmospheres, which expand and contract in regular cycles that typically last between 1 and 100 days. As CCs expand, their brightness increases by more than a factor of 10 (up to two magnitudes), before beginning to fade again as the star shrinks.

Online resources: the magnitude system

You may not be familiar with the concept of defining the brightness of astrophysical objects using the magnitude system. In that case, the online resources for this chapter provide some additional material from a Stage 2 astronomy module that you can refer to.

In 1907, the American astronomer Henrietta Swan Leavitt observed a large number of CCs in the LMC and discovered that the periods of CCs' regular brightness variations are correlated with their mean *luminosities*.

- How do you think Henrietta Swan Leavitt was able to determine the intrinsic luminosities of the CCs she observed?
- She was able to measure their distances using stellar parallax! In Exercise 5.3 you computed the stellar parallax of a star in the LMC. This *is* a small angle, but it *was* measurable using the technology that was available in 1907. Once the distances to the CCs were known, she could use their observed brightnesses to compute their luminosities.

The **period–luminosity relation** discovered by Leavitt has since been corroborated by numerous observations using more modern telescopes, including the *Hubble Space Telescope (HST)*. The relation is normally expressed in terms of the mean **absolute magnitude** $\langle M \rangle$ of the star, the period P of the brightness variation in *days*, and two empirically determined parameters, A and B .

Classical Cepheid period–luminosity relation

$$\langle M \rangle = A [\log_{10}(P) - 1] - B \quad (5.24)$$

Equation 5.24 defines a linear relationship between $\langle M \rangle$ and $[\log_{10}(P) - 1]$, and B represents the intercept value of $\langle M \rangle$ when $[\log_{10}(P) - 1] = 0$.

Physically, this means that B is the mean absolute magnitude of a ‘standard’ CC that has a period of exactly 10 days. Using *HST* observations, astronomers have derived values for $A = -2.43 \pm 0.12$ and $B = 4.05 \pm 0.02$ using V-band optical observations of nine CCs in the Milky Way (Benedict *et al.*, 2007).

Equation 5.24 allows cosmologists to infer the intrinsic brightness of a CC by measuring the period of its variability cycle. In the next example you will see how to use this information to work out how far away the CC is, without needing to be able to measure its parallax.

- When cosmologists use the period–luminosity relation to estimate the distance to CCs, which of the distance definitions discussed in Section 5.1 are they actually measuring?
- They are estimating the distance to the CCs using their apparent brightness, so they are measuring the *luminosity distance*, d_L .

Example 5.4

Using *HST* observations, astronomers have measured the period and mean V-band apparent magnitude of δ Cephei to be $P = 5.36627$ days and $\langle m_V \rangle = 3.960$ magnitudes, respectively. Use this information and the A and B parameters derived by Benedict *et al.* (2007) to compute the distance to δ Cephei in parsecs.

Solution

The difference between the apparent magnitude m of an object and its absolute magnitude M is related to the object's luminosity distance d_L in parsecs by a quantity called the **distance modulus**, which is typically represented using the symbol μ .

Distance modulus

$$\mu = m - M = 5 \log_{10}(d_L/\text{pc}) - 5 \quad (5.25)$$

For the specific case of the V band, we can write:

$$m_V - M_V = 5 \log_{10}(d_L) - 5$$

We substitute for M_V using Equation 5.24:

$$\langle m_V \rangle - A [\log_{10}(P) - 1] + B = 5 \log_{10}(d_L) - 5$$

We have been given the period in days and we are told that our mean magnitude measurement is for the V band, so we can assume that $A = -2.43$ and $B = 4.05$ using the result from Benedict *et al.* (2007).

Now it is straightforward to compute:

$$\begin{aligned} \log_{10} d_L &= \frac{5 + m_V - A [\log_{10}(P) - 1] + B}{5} \\ &= \frac{5 + 3.960 + 2.43 \times [\log_{10}(5.36627) - 1] + 4.05}{5} \\ &= 2.47 \\ \implies d_L &= 295 \text{ pc} \end{aligned}$$

The brightness of classical Cepheids means that they can be detected at much larger distances than those that can be measured using stellar parallax. The unprecedented sensitivity of the *HST* allowed Newman *et al.* (1999) to detect CCs and measure their periods in the galaxy NGC 4603, which is 33 Mpc away!

- What is the redshift of NGC 4603? How fast is it receding due to cosmic expansion, and how does this value compare with the peculiar velocity of the Milky Way ($\sim 600 \text{ km s}^{-1}$)?
- According to the Hubble–Lemaître law (Equation 1.2) the redshift of NGC 4603 is $z \approx H_0 D/c \approx 0.007$ for $D = 33 \text{ Mpc}$. This corresponds to a cosmological recession velocity $v_r \approx cz \approx 2230 \text{ km s}^{-1}$. This is approximately four times faster than the peculiar velocity of the Milky Way, which means that the influence of cosmic expansion is clearly detectable, albeit not completely dominant, at this distance.



Figure 5.7 A Type Ia supernova (SN 1994D) in the galaxy NGC 4526. The supernova is the bright star-like object in the bottom left. For a short time it outshines all other stars in the galaxy combined!

Type Ia supernovae

Earlier in Section 5.2.2 you read that classical Cepheids are only bright enough to enable measurement of the luminosity distance to galaxies that are ~ 30 Mpc away. At this distance galaxies' peculiar velocities are comparable to their cosmological recession velocities. To measure the luminosity distances to objects with motions that are *dominated* by the Hubble flow, cosmologists must find standard candles that are thousands of times brighter than CCs.

The vast majority of stars end their lives as **white dwarfs** (WDs). WDs are the remnants of stellar cores after all nuclear fusion reactions within them have stopped. They are extremely dense, with masses in the range $\sim 0.1 M_{\odot}$ to $1 M_{\odot}$ compressed into a sphere roughly the size of the Earth!

The luminosities of WDs are typically $< 10^{-2} L_{\odot}$ and they vary over a large range, so they cannot be used by themselves as standard candles. However, sometimes a WD forms in a binary system with another star that is still burning nuclear fuel. In some of these systems, the WD starts to accrete gas and plasma from its binary companion and its mass slowly increases.

Eventually, the mass of the WD approaches a theoretical threshold called the **Chandrasekhar limit** $M_C \approx 1.39 M_{\odot}$, above which the star's internal pressure cannot counteract the force of its self-gravity. Just before it gets to this limit, carbon fusion ignites in its outer layers and starts a runaway thermonuclear reaction that completely destroys the remnant. This detonation is called a **Type Ia supernova** (Figure 5.7), and releases so much energy ($\sim 10^{44}$ J) that the exploding WD briefly outshines all of the stars in its host galaxy combined. This astonishing luminosity means that Type Ia supernovae can be detected at very large distances.

The scenario we just described – with a single WD accreting material from a stellar companion – is actually quite rare. Only around 20% of Type Ia supernovae are believed to originate this way. More frequently, the stellar companion itself also runs out of fuel, and a binary WD system forms. Subsequent tidal interactions between the two WDs slowly remove angular momentum from the system, and the stellar remnants orbit closer and closer to each other. Eventually, the orbital separation becomes so small that the more massive WD is able to tidally disrupt its smaller companion and accrete material from it. The accretion happens much more rapidly than it did in the single WD scenario, but the ultimate end state is very similar. As the more massive WD approaches the Chandrasekhar limit, a thermonuclear detonation is ignited and the remnant is destroyed, forming a Type Ia supernova.

Regardless of their origin, the fact that all Type Ia supernovae occur when accreting white dwarfs reach a *specific* mass means that the amount of energy they release on detonation is almost identical. Theoretically, this means that the peak luminosity of every Type Ia supernova in the Universe is almost exactly the same, making them excellent standard candles.

In reality the situation is slightly more complicated. The actual peak luminosity depends somewhat on the details of the WD's exact mass and composition at the time of the explosion, and the immediate environment surrounding the supernova can also modify and attenuate the radiation that we eventually observe. Fortunately, astronomers have discovered that the shapes of the supernova's **light curves** – time-ordered series of measurements of the object's apparent brightness obtained at different times – when measured in different wavelength bands can be used to infer how these different effects have impacted the maximum brightness that the supernova actually reached.

To derive a quantitative relationship between light curve shapes and intrinsic luminosities of Type Ia supernovae in distant galaxies, cosmologists use an approach that is conceptually similar to the one used to calibrate the Cepheid period–luminosity relation. By measuring the periods and inferring the luminosities of CCs in the same galaxies as the supernovae, they can work out how far away those supernovae really are. Then, by observing the apparent brightnesses of the supernovae, cosmologists can infer how intrinsically luminous they are. Finally, they can measure the light curve shapes for the different supernovae and calibrate a brightness–shape relation. CCs are only bright enough to be detectable at relatively low redshifts (33 Mpc corresponds to $z \sim 0.01$), so cosmologists must assume that the supernovae in the local Universe behave in the same way as their more distant counterparts.

- Why is this assumption about the similarity of Type Ia supernovae throughout the Universe valid?
- This is a good assumption because the behaviour of Type Ia supernovae is governed by the basic laws of physics that govern star formation, evolution and destruction. We do not expect these laws to vary according to spatial location or as the Universe ages.

The light curves in Figure 5.8a show how supernovae that are more intrinsically luminous also tend to fade more slowly over time. While these curves all look generally similar in shape, their peak luminosities are clearly not similar enough for them to be used as standard candles.

To use observed Type Ia supernovae as standard candles, cosmologists define an idealised ‘standard’ supernova with a specific light curve shape and corresponding intrinsic luminosity (Figure 5.8b). The observed peak magnitudes of real supernovae are ‘corrected’ to represent the peak apparent magnitude that *would* have been observed for the standard supernovae, if it was placed at the same luminosity distances as the real ones. To reflect the fact that differences in their intrinsic brightness must be corrected in order to use their light curve shapes, Type Ia supernovae are often referred to as ‘*standardisable* candles’.

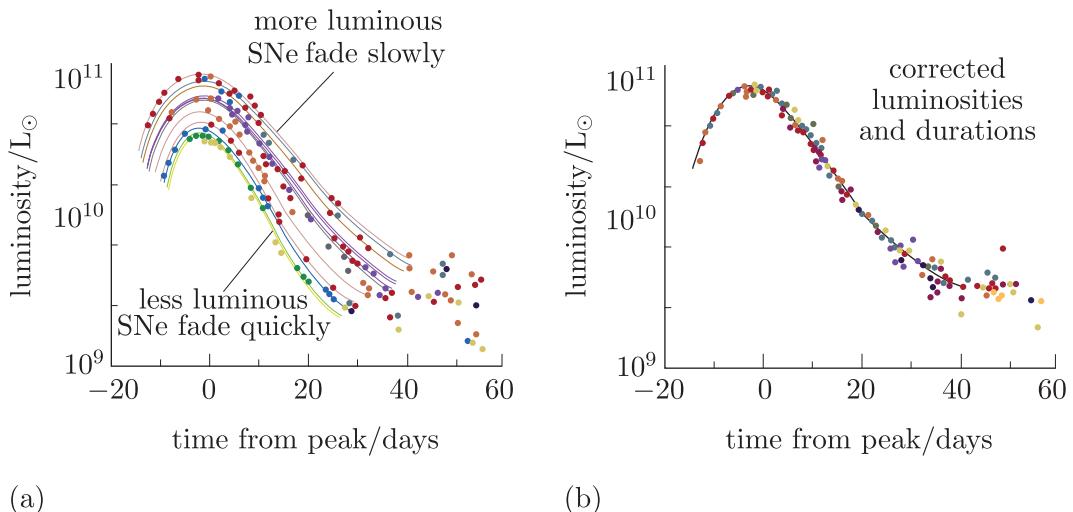


Figure 5.8 (a) Selection of observed Type Ia supernovae (or ‘SNe’) light curves, showing that more luminous objects tend to fade more slowly than less luminous ones. (b) The same light curves after standardisation, where the data have been calibrated so that the curves appear to lie on top of each other.

Type Ia supernova are very rare events. We would expect approximately one to occur per century in a galaxy like the Milky Way.* Nonetheless, there so many galaxies in the Universe that astronomers have detected thousands of Type Ia supernovae during the last 80 years. As the sensitivity of telescopes has improved, supernovae have been discovered at larger and larger distances. At the time of writing (2023), the most distant Type Ia supernova that has been observed is SN UDS10Wil, which has a measured redshift $z = 1.914$.

The 2011 Nobel Prize in Physics was awarded to astronomers from two independent teams – the Supernova Cosmology Project (SCP) and High-z Supernova Search Team – who used observations of Type Ia supernovae at high redshift to demonstrate a completely unexpected result.

They discovered that that $\Omega_\Lambda > 0$, and the expansion of the Universe is accelerating. The next section describes how this discovery was made.

5.3 Measuring H_0 and density parameters

In Chapter 4 you saw that the Friedmann equations and the values of H_0 , $\Omega_{\text{m},0}$, $\Omega_{\text{r},0}$, $\Omega_{\Lambda,0}$ and k completely determine the expansion history of the Universe, its future evolution and its ultimate fate.

Measuring these parameters allows cosmologists to determine how physical properties of the Universe – like temperature and density – have evolved in

*The closest recent Type Ia supernova was observed in the Large Magellanic Cloud in 1987. It was named SN 1987A to reflect the fact that it was the first Type Ia supernova to be observed that year.

the past, and to predict how they will continue to evolve in the future. Their values also tell cosmologists about the fundamental properties of the Universe like its spatial curvature, its age and its physical scale.

So far in this chapter we have introduced the concepts and some of the techniques that cosmologists use in order to constrain these values using observational data. Next we will see how these techniques are actually used in practice. You will learn that, by measuring d_L for large numbers of distant objects, cosmologists are able to *constrain* H_0 and two of the four density parameters.

This is not as simple as it might sound. Figures 5.2b and 5.4b, in Sections 5.1.2 and 5.1.3 respectively, show that the predicted values of d_L (and d_A) for $z \lesssim 0.1$ are very similar for model universes with very different cosmological parameters. To precisely constrain the values of the cosmological parameters of the real Universe, cosmologists need to distinguish between potential models that are much more similar than those illustrated in Figures 5.2 and 5.4. Their goal is to precisely measure contributions of matter, radiation and Λ to the energy density of the Universe. This means that they need very precise measurements of d_L (or d_A) for galaxies at redshifts $z \gg 0.1$, which is a technical challenge that has only recently been overcome.

5.3.1 Measuring the Hubble constant

There are many reasons that cosmologists would *like* to know the value of H_0 , beyond the simple fact that it measures the current expansion rate of the Universe. For example, it can be used to define a **Hubble distance**, d_H . This common cosmological unit has dimensions of length, and sets an approximate size scale for the observable Universe.

The Hubble distance

$$d_H = c/H_0 \quad (5.26)$$

Cosmologists also define a quantity called the **Hubble time**, t_H , which provides an approximate time scale for the age of the Universe.

Note that in general the Hubble time and the true age of the Universe are *not* equal.

The Hubble time

$$t_H = H_0^{-1} \quad (5.27)$$

Later in this chapter you will see that the *overall* matter density of the Universe can be constrained without knowing H_0 . However, for reasons that we will touch on in a later chapter, the fractions of baryons and dark

matter can often only be *measured* as multiples of a quantity denoted as h^2 and defined as

$$h^2 = \left(\frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}} \right)^2 \quad (5.28)$$

This means that measuring the density of baryons in the Universe requires knowledge of the true value of H_0 .

Exercise 5.4

Estimate a value for the Hubble time in units of My (millions of years); use the central value of H_0 listed in the table of constants.

One way to estimate the Hubble constant is to *assume* a cosmological model for the Universe and simultaneously find ranges of that model's parameters – like H_0 and the density parameters – that are consistent with actual observations. In fact, that is how the value of H_0 provided in this book was derived. That value *assumes* a flat Universe with $\Omega_k = 0$ and that Λ truly is a cosmological constant, with equation of state parameter $w_\Lambda = -1$. Neither of these assumptions is derived from fundamental physical principles, and neither has to be true.

It would be much better if we could find some way of measuring H_0 that is completely independent of any assumptions about the Universe or the values of the other cosmological parameters. As we discussed earlier this chapter, it should be possible to infer H_0 directly if we can measure the proper distances and redshifts for a large sample of objects whose motions are dominated by the Hubble flow.

We have identified Type Ia supernovae as a population of standardisable candles that are bright enough to be detectable in galaxies as far away as $z \sim 1.5$. It turns out that if we know the *true* peak absolute magnitude M_{true} of a ‘standard’ Type Ia supernova, then we can use the standardised (‘corrected’) apparent peak magnitudes m_{corr} of many distant supernovae to directly measure H_0 . The following exercise asks you to derive an expression that will be very useful for the remainder of this section, and again in Section 5.3.3.

Exercise 5.5

Use the definition of the distance modulus (Equation 5.25) to show that the brightest apparent magnitude m of a Type Ia supernova can be expressed as:

$$m = M - 5 \log_{10} H_0 + 5 \log_{10} (H_0 d_L) + 25 \quad (5.29)$$

where M is the supernova’s brightest absolute magnitude, d_L is the luminosity distance to the supernova in Mpc and H_0 is the present Hubble constant expressed in $\text{km s}^{-1} \text{ Mpc}^{-1}$.

By rearranging the definition of the distance modulus in Equation 5.29 we can write

$$\log_{10} H_0 = \log_{10}(H_0 d_L) - \frac{m_{\text{corr}}}{5} + \frac{M_{\text{true}}}{5} + 5 \quad (5.30)$$

For $z \ll 1$ we know $H_0 d_L = H_0(1+z)d_p(t_0) \approx cz$, and we can define the variable

$$\mathcal{A} = \log_{10}(H_0 d_L) - \frac{m_{\text{corr}}}{5} \approx \log_{10}(cz) - \frac{m_{\text{corr}}}{5} \quad (5.31)$$

The approximation $H_0 d_L \approx cz$ holds at low redshifts, and the value of \mathcal{A} can be directly measured using observations of a sample of Type Ia supernovae. It is just the intercept of a linear fit to the logarithms of the supernova redshifts and corrected apparent peak magnitudes. However, at high redshifts the situation is more complicated. Exercise 5.6 explores how to estimate the redshift at which the values $H_0 d_L$ and cz become significantly different.

Exercise 5.6

The Hubble–Lemaître law, $H_0 d_L \approx cz$, provides a good approximation for the relationship between the measured redshifts and luminosity distances of nearby objects, for which $z \ll 1$. However, for more distant objects, this approximation becomes increasingly inaccurate. An alternative expression, which is accurate to within 1% for values of $z < 0.2$, is:

$$H_0 d_L = cz \left(1 + \frac{1 - q_0}{2} z \right) \quad (5.32)$$

where q_0 is the deceleration parameter from Equation 4.43.

Assuming that the Universe is flat and that $\Omega_{r,0} \approx 0$, use the cosmological parameters listed in the table of constants to show that by $z = 0.13$, the result of Equation 5.32 is already 10% greater than the $H_0 d_L \approx cz$ approximation that applies for $z \ll 1$.

Once \mathcal{A} has been measured, we just need to find the true peak absolute magnitude (M_{true}) of a Type Ia supernova, and then we can use Equation 5.30 to compute H_0 directly. In the next section you will learn how cosmologists construct a cosmological distance ladder that ultimately allows M_{true} to be determined.

5.3.2 Cosmological distance ladders

To determine the value of M needed to compute H_0 in Equation 5.30, cosmologists need to measure the corrected apparent magnitude (M_{corr}) of at least one Type Ia supernova for which the proper distance is already known. If a Type Ia supernova exploded in the Milky Way we could measure its distance directly using parallax, and immediately compute its absolute magnitude. If astronomers could accurately measure the

supernova's light curve then we could also compute what its standardised absolute magnitude would be.

However, as mentioned in Section 5.2.2, Type Ia supernovae are very rare events: with an occurrence rate of roughly one per decade within 20 Mpc of the Sun, we would be very lucky to detect a single instance within our galaxy within the next century. Even if we did observe one, that supernova might be unusual in some way and then our single measurement of M would be misleading.

To accurately measure the distances to large numbers of Type Ia supernovae, cosmologists start by measuring the distances to fainter standard candles like classical Cepheids, which exist in all galaxies and are therefore much more likely to be found nearby. The goal is to find examples of Type Ia supernovae in galaxies that also contain detectable CCs. The distances to those CCs, and by extension their host galaxies, can then be inferred using the period–luminosity relation. The following exercise asks you to estimate the rate at which Type Ia supernovae occur in galaxies that are close enough for the CCs they contain to be detected.

Exercise 5.7

In Section 5.2.2 you read that CCs can be detected out to redshifts $z \sim 0.01$. Using the value of H_0 listed in the table of constants, calculate how many Type Ia supernovae would you expect to find per year in galaxies that contain *detectable* CCs?

The distances to galaxies containing observable CCs may be relatively small on the scale of the observable Universe, but they are still very large when compared to the size of a single galaxy (typically ~ 10 kpc).

Therefore, we can assume that the distances to the CCs and a Type Ia supernova within a single galaxy are effectively identical. If we could directly measure the distance to the CCs, then we would also know the true distance to the Type Ia supernovae in their host galaxies. With enough examples, we could then straightforwardly calibrate the light-curve-based standardisation procedure that we need in order to accurately infer the distances to supernovae at much larger redshifts.

However, as you read in Section 5.2.2, the distances to CCs in distant galaxies cannot be measured directly either: they must be inferred from their apparent magnitudes and their variability periods using Leavitt's period–luminosity relation. To calibrate this relation, astronomers need to find CCs that are close enough for their distances to be measured directly using parallax.

Figure 5.9 shows how cosmologists measure the value of H_0 by constructing a **cosmological distance ladder**. To 'anchor' the distance ladder, distances measured using stellar parallax for CCs in the Milky Way and its satellite galaxies are used to calibrate the CC period–luminosity

relation (Equation 5.24). The next rung on the distance ladder uses more-distant CCs to calibrate the peak absolute magnitude M of a ‘standard’ Type Ia supernova. Finally, by assuming this value for M and measuring the apparent magnitudes of more distant supernovae, cosmologists use Equation 5.30 to measure the value of H_0 . In 2022, cosmologists using this distance ladder method reported a directly measured value of H_0 to be $73.04 \pm 1.04 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

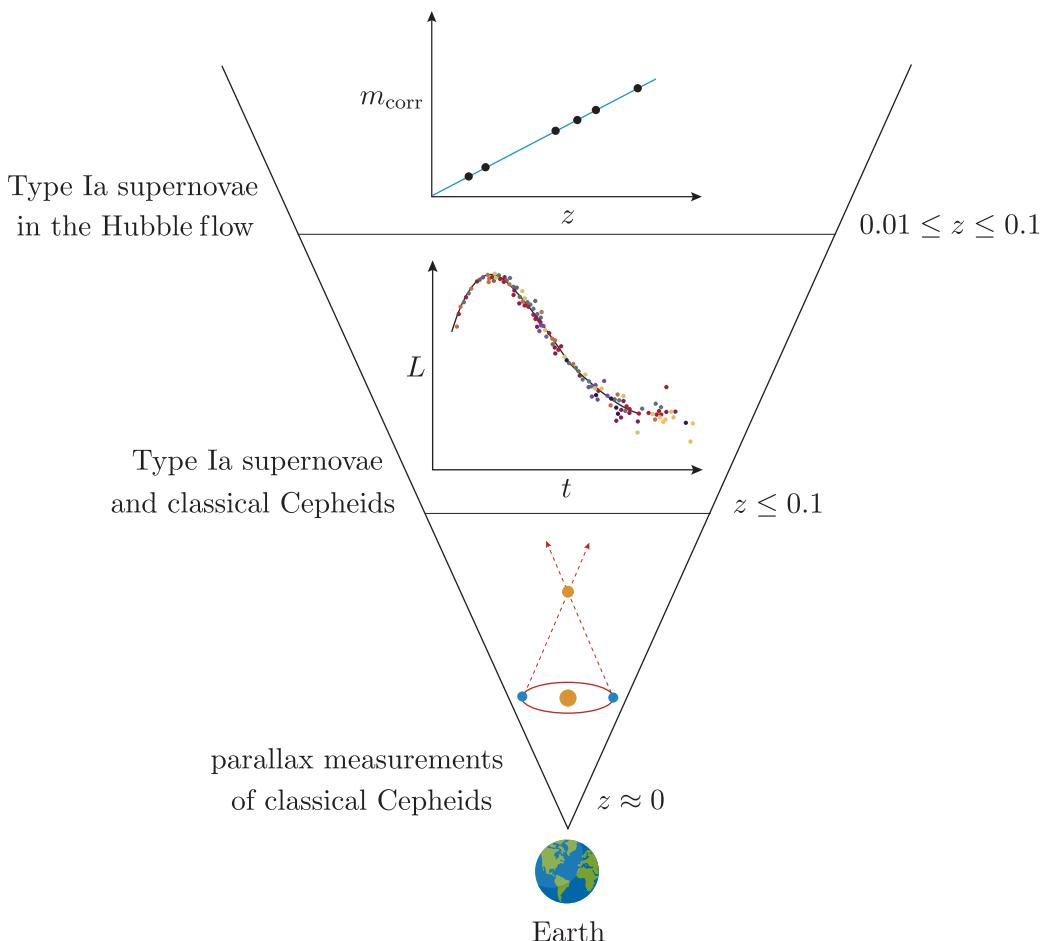


Figure 5.9 A schematic of the cosmological distance ladder. The ladder is anchored by measuring distances to nearby CCs using stellar parallax. The distances to these anchored CCs are used to calibrate the period–luminosity relation Equation 5.24 which allows their intrinsic luminosities to be inferred from their periodic variability timescales. On the second rung of the ladder CCs are used to measure the distances to distant galaxies that host Type Ia supernovae. The light curves and apparent magnitudes of these supernovae are used to calibrate the standardisation procedure that allows them to be used as standard candles. On the third rung, a sample of more distant Type Ia supernovae whose motions are dominated by the Hubble flow are identified. Their apparent magnitudes are standardised and used together with Equation 5.30 to measure H_0 .

5.3.3 Measuring the density parameters for matter and Λ

In this section you will learn how observations of distant Type Ia supernovae can also be used to constrain the present-day density parameters of matter and Λ . For simplicity, we will assume a flat universe in which Ω_r is negligible, but the techniques being explored can also be used to constrain more complicated models for the Universe.

We can start by noting that the version of the distance modulus expression shown in Equation 5.29 is just the equation of a straight line. It expresses a linear relationship between m and $\log_{10} H_0 d_L$ with intercept \mathcal{M} , which can be written as:

$$\mathcal{M} \equiv M - 5 \log_{10} H_0 + 25 \quad (5.33)$$

At low redshift, \mathcal{M} can be measured if we can identify a sample of Type Ia supernovae that are close enough to use the approximation $d_L \approx cz/H_0$. For any one of these supernovae, we could rewrite Equation 5.29 as

$$m = M - 5 \log_{10} H_0 + 5 \log_{10}(cz) + 25 = \mathcal{M} + 5 \log_{10}(cz) \quad (5.34)$$

To measure d_L in Equation 5.29 we would need to know the intrinsic peak luminosity of the Type Ia supernovae in our sample but, unlike d_L , the quantity z in Equation 5.34 is *directly* measurable. To estimate \mathcal{M} , all we need to do is fit a straight line to measured values of m and $\log_{10}(cz)$ and read off the intercept.

Now that we know how to calculate \mathcal{M} , we could solve Equation 5.29 to determine d_L for a sample of high-redshift supernovae without needing to know their absolute magnitudes. We just need to assume that the low- and high-redshift supernovae have the *same* absolute magnitude so that \mathcal{M} is equal for both samples. In Section 5.2.2 you read that the peak apparent magnitudes of Type Ia supernovae can be corrected based on their light curve shapes to represent a ‘standard’ Type Ia supernova with a *specific* absolute magnitude. Therefore, if we use corrected apparent magnitudes in Equation 5.29, then we *can* assume that \mathcal{M} is completely independent of redshift. You will apply these ideas in Examples 5.5 and 5.6.

Example 5.5

In this example you will perform calculations that will let you determine whether two small sets of real observational data are consistent with a flat, matter-only model for the Universe.

Tables 5.1 and 5.2 list the redshifts and corrected apparent magnitudes for two samples of Type Ia supernovae. Values in the ‘ σ_z ’ and ‘ $\sigma_{m_{\text{corr}}}$ ’ columns represent the 1σ uncertainty on these properties. Use the tables as directed to answer the following questions.

Table 5.1 Redshifts (z) and corrected apparent peak magnitudes (m_{corr}) for a sample of low-redshift supernovae, in order of date observed, as measured by the Calán/Tololo Supernova Survey.

Supernova	z	σ_z	m_{corr}	$\sigma_{m_{\text{corr}}}$
1990O	0.030	0.002	16.26	0.20
1992ae	0.075	0.002	18.43	0.20
1992aq	0.101	0.002	19.16	0.23
1992bo	0.018	0.002	15.61	0.21
1992bp	0.079	0.002	18.27	0.18

(Adapted from Hamuy *et al.*, 1993)

Table 5.2 Redshifts (z) and corrected apparent peak magnitudes (m_{corr}) for a sample of high-redshift supernovae, in order of date observed, as measured by the Supernova Cosmology Project.

Supernova	z	σ_z	m_{corr}	$\sigma_{m_{\text{corr}}}$
1995aw	0.400	0.030	22.36	0.19
1995ax	0.615	0.001	23.19	0.25
1997L	0.550	0.010	23.51	0.25
1997S	0.612	0.001	23.69	0.21
1997aj	0.581	0.001	23.09	0.22

(Adapted from Perlmutter *et al.*, 1997)

- Estimate \mathcal{M} by using the low-redshift data in Table 5.1 to evaluate Equation 5.33 for each supernova, and then averaging the results.
- Use your estimate for \mathcal{M} to calculate the luminosity distance to each of the high-redshift supernovae listed in Table 5.2, in units of the Hubble distance, $d_{\text{H}} = c/H_0$.
- Now, use Equation 5.9 and Equation 5.12 to compute a *theoretical* prediction for the luminosity distance, in d_{H} units, to each high-redshift supernova, assuming $k = 0$ and $\Omega_{\text{m}} = 1$.

Solution

- If we label the five low-redshift supernovae in Table 5.1 with an index $i = 1, 2, \dots, 5$, then we can compute \mathcal{M}_i for each supernova by rearranging equation Equation 5.34.

$$\mathcal{M}_i = m_{\text{corr},i} - 5 \log_{10} z_i - 5 \log_{10} c$$

Table 5.3 lists the result of this computation for each supernova. Note that the numerical results throughout this example are calculated using c in units of kilometres per second.

Table 5.3 Computed results for the low-redshift supernova sample.

Supernova	\mathcal{M}_i
1990O	-3.510
1992ae	-3.329
1992aq	-3.246
1992bo	-3.051
1992bp	-3.602

Chapter 5 Measuring cosmological parameters

Now we can combine the results for the individual supernovae and compute a mean value for \mathcal{M} to use as our final estimate:

$$\mathcal{M} = \langle \mathcal{M}_i \rangle = \frac{1}{5} \sum_{i=1}^5 \mathcal{M}_i = -3.348$$

- (b) For brevity will use the symbol \mathcal{D} to denote d_L in units of d_H . To calculate \mathcal{D} for the high-redshift supernovae listed in Table 5.2 we can label each of objects with a index $j = 1, 2, \dots, 5$, then rearrange Equation 5.29 to give:

$$\mathcal{D}_j = \frac{d_{L,j}}{d_H} = \frac{d_{L,j} H_0}{c} = \frac{1}{c} 10^{\alpha_j}$$

where we have defined

$$\alpha_j = \frac{m_{\text{corr},j} - \mathcal{M}}{5}$$

Table 5.4 lists computed values of \mathcal{D} for the high-redshift sample.

Table 5.4 Computed values of \mathcal{D} for the high-redshift supernovae.

Supernova	\mathcal{D}_j
1995aw	0.462
1995ax	0.677
1997L	0.785
1997S	0.852
1997aj	0.647

- (c) To determine whether these results are consistent with a flat matter-only universe with $\Omega_m = 1$ we need to compare them with a theoretical prediction that we can compute using Equation 5.9 and Equation 5.12. Combining these equations, assuming that $\Omega_r = \Omega_\Lambda = \Omega_k = 0$ and $\Omega_m = 1$, we can write

$$\begin{aligned} \mathcal{D}_{j,\text{predicted}} &= (1 + z_j) \int_0^{z_j} \frac{dz'}{\sqrt{(1 + z')^3}} \\ &= (1 + z_j) \left[-\frac{2(1 + z')}{\sqrt{(1 + z')^3}} \right]_0^{z_j} \\ &= 2(1 + z_j) \left[1 - \frac{(1 + z_j)}{\sqrt{(1 + z_j)^3}} \right] \end{aligned}$$

Table 5.5 lists the predicted values of \mathcal{D} for the high-redshift sample.

Table 5.5 Computed and predicted values of \mathcal{D} for the high-redshift supernovae.

Supernova	\mathcal{D}_j	$\mathcal{D}_{j,\text{predicted}}$
1995aw	0.462	0.434
1995ax	0.677	0.688
1997L	0.785	0.610
1997S	0.852	0.685
1997aj	0.647	0.647

To determine whether the observational data are consistent with the corresponding theoretical predictions, it is important to correctly account for the measurement uncertainties. To do this you will need to use two standard results related to the calculation and propagation of experimental uncertainties.

First, if the quantity x is measured with an associated uncertainty σ_x^2 , then the result of applying a unary function f to x is also uncertain, such that:

$$\sigma_{f(x)}^2 = \left(\frac{df(x)}{dx} \right)^2 \sigma_x^2 \quad (5.35)$$

Second, if a and b are two *uncorrelated* measurements with uncertainties σ_a^2 and σ_b^2 , respectively, then the result of applying a binary function f to a and b is also uncertain, such that:

$$\sigma_{f(a,b)}^2 = \left(\frac{\partial f(a,b)}{\partial a} \right)^2 \sigma_a^2 + \left(\frac{\partial f(a,b)}{\partial b} \right)^2 \sigma_b^2 \quad (5.36)$$

Here we have used the standard notation $\partial y / \partial x$ to denote the partial derivative of y with respect to x .

The following example shows how you can use these two results to estimate the uncertainties associated with the quantities you calculated in Example 5.5. Note that this example reuses some of the notation and results from Example 5.5.

A unary function is one that operates on a single variable.

A binary function is one that operates on two variables.

Example 5.6

Use standard techniques for propagation of measurement uncertainties and the values listed in Tables 5.1 and 5.2 to do the following.

- (a) Propagate the observational uncertainties listed in Table 5.1 to calculate the 1σ uncertainty in your estimate for \mathcal{M} from Example 5.5.
- (b) Use this estimate for the uncertainty in \mathcal{M} and the observational uncertainties listed in Table 5.2 to calculate 1σ uncertainties for each of your estimated luminosity distances from part (b) of Example 5.5.

Solution

- (a) To estimate the uncertainty on \mathcal{M}_i for each separate supernova we can use Equation 5.35 (with $f = \log_{10}$) to propagate the 1σ errors listed in Table 5.1.

$$\begin{aligned} \sigma_{\mathcal{M}_i}^2 &= \sigma_{m_{\text{corr},i}}^2 + 5 \sigma_{z_i}^2 \left[\frac{d}{dz_i} \log_{10} z_i \right]^2 \\ &= \sigma_{m_{\text{corr},i}}^2 + 5 \left[\frac{\sigma_{z_i}}{z_i \ln 10} \right]^2 \end{aligned}$$

The results of this calculation for each low-redshift supernova are listed in Table 5.6.

Table 5.6 Computed uncertainties and their squares for the low-redshift supernova sample.

Supernova	$\sigma_{\mathcal{M}_i}$	$\sigma_{\mathcal{M}_i}^2$
1990O	0.210	0.0441
1992ae	0.202	0.0408
1992aq	0.231	0.0534
1992bo	0.236	0.0557
1992bp	0.182	0.0331

Summing the individual results gives the overall uncertainty on \mathcal{M} :

$$\sigma_{\mathcal{M}}^2 = \frac{1}{5} \sum_{i=1}^5 \sigma_{\mathcal{M}_i}^2 = 0.045$$

The final value of $\sigma_{\mathcal{M}}$ is calculated by taking the square root of $\sigma_{\mathcal{M}}^2$:

$$\sigma_{\mathcal{M}} = 0.213$$

- (b) We can assume that $m_{\text{corr},j}$ and \mathcal{M} are uncorrelated, so the uncertainties for each luminosity distance estimate \mathcal{D}_j can be computed using Equation 5.36 repeatedly (with $f = 10^g$ and $g = \alpha_j$).

$$\sigma_{\mathcal{D}_j}^2 = \sigma_{m_{\text{corr},j}}^2 \left[\frac{d}{dm_{\text{corr},j}} \frac{10^{\alpha_j}}{c} \right]^2 + \sigma_{\mathcal{M}}^2 \left[\frac{d}{d\mathcal{M}} \frac{10^{\alpha_j}}{c} \right]^2 \quad (5.37)$$

The derivatives can be evaluated using the chain rule:

$$\frac{d}{dm_{\text{corr},j}} 10^{\alpha_j} = \frac{d}{d\alpha_j} 10^{\alpha_j} \cdot \frac{d\alpha_j}{dm_{\text{corr},j}} = \ln(10) 10^{\alpha_j} \cdot \frac{1}{5}$$

and

$$\frac{d}{d\mathcal{M}} 10^{\alpha_j} = \frac{d}{d\alpha_j} 10^{\alpha_j} \cdot \frac{d\alpha_j}{d\mathcal{M}} = \ln(10) 10^{\alpha_j} \cdot -\frac{1}{5}$$

Substituting these two results into Equation 5.37, our uncertainty estimates for \mathcal{D} become

$$\sigma_{\mathcal{D}_j}^2 = \frac{1}{25c^2} \left\{ \sigma_{m_{\text{corr},j}}^2 [\ln(10) \cdot 10^{\alpha_j}]^2 + \sigma_{\mathcal{M}}^2 [\ln(10) \cdot 10^{\alpha_j}]^2 \right\}$$

Table 5.7 lists computed values of $\sigma_{\mathcal{D},j}$ for the high-redshift sample.

Table 5.7 Computed uncertainties on the luminosity distance for the high-redshift supernova sample.

Supernova	$\sigma_{\mathcal{D}}$
1995aw	0.061
1995ax	0.102
1997L	0.119
1997S	0.117
1997aj	0.091

Tables of data like Tables 5.4 and 5.7 can be difficult to interpret. Instead, look at Figure 5.10, which presents the results from Examples 5.5 and 5.6 in graphical form.

Now we can straightforwardly observe that three of our \mathcal{D} estimates *are* consistent with a flat, matter-only universe. However, one of the other estimates is consistent with a very different universe that only contains a cosmological constant, and another appears to suggest an intermediate model with matter and a Λ component. This example illustrates just how difficult it is to constrain the cosmological parameters once experimental uncertainties are taken into account.

The approach taken in Example 5.5 is very similar to the way that professional cosmologists use Type Ia supernova observations to constrain the cosmological parameters. They also use a low-redshift sample of supernovae to constrain \mathcal{M} , and compare their results for high-redshift supernovae with model predictions, to determine what combinations of cosmological parameters are consistent with their observations. The differences lie primarily in the sophistication of the statistical techniques that are used, and the number of different supernovae that are considered.

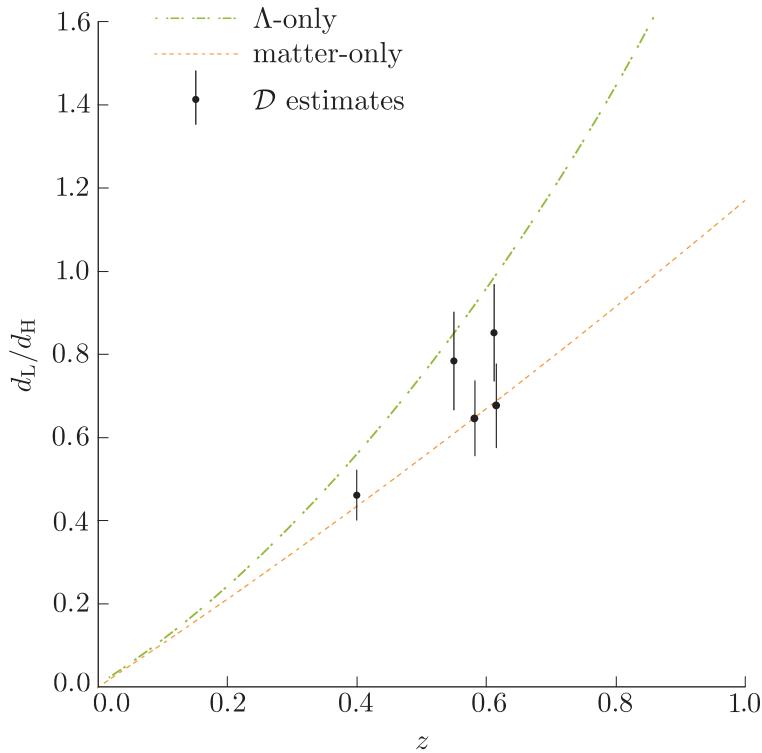


Figure 5.10 Estimates of \mathcal{D} and its uncertainty for the high-redshift supernova sample referenced in Tables 5.2 and 5.4. Predicted curves of $\mathcal{D}(z)$ are shown for flat universes that contain only matter (orange dot-dashes) or only a cosmological constant (short green dashes). Different subsets of the estimated supernova \mathcal{D} values are consistent with both universe models at the 1σ level.

A noteworthy feature of this method is that we can constrain plausible values for Ω_m and Ω_Λ without needing to know the value of H_0 . The expression for \mathcal{M} contains H_0 , but we were able to measure \mathcal{M} directly using low-redshift supernovae so the value of H_0 was not required.

Similarly, using Equations 5.9 and 5.12 allows the composite quantity $H_0 d_L$ to be predicted directly given values of the density parameters. In Section 5.3.1 you saw that measuring H_0 is complicated, and requires the true luminosities of standard candles to be known. The ability to measure other cosmological parameters independently is therefore a big advantage!

Figure 5.11 shows the Hubble diagram for 60 Type Ia supernovae that were analysed by the Supernova Cosmology Project (SCP) team. The supernova magnitudes you used in Example 5.5 are actually a subset of these data. The curves in Figure 5.11 show predictions for how the corrected apparent peak magnitudes of Type Ia supernovae should evolve as a function of redshift for different cosmological models. It might not be obvious under visual inspection, but a careful statistical analysis by the SCP team showed that the high-redshift data points ($z \gtrsim 0.2$) are *inconsistent* with a universe in which Ω_Λ is zero. This was the historic discovery for which the SCP team jointly won the 2011 Nobel Prize in Physics.

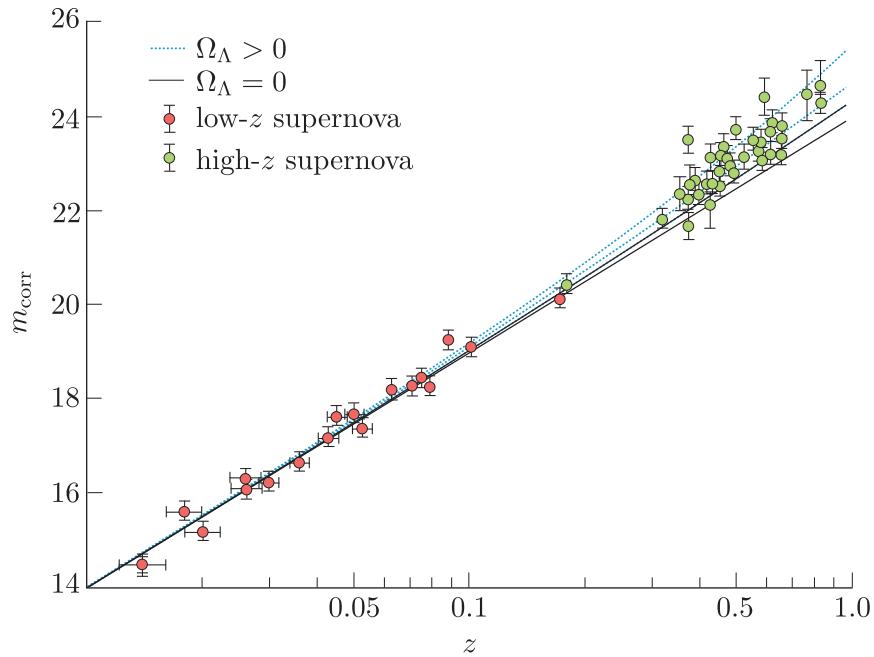


Figure 5.11 A Hubble diagram showing the corrected peak magnitudes of 18 low-redshift supernovae (red points) and 42 high-redshift supernovae (green points) versus their redshifts (after Perlmutter *et al.*, 1999). The different curves show predictions for different model universes.

Our constraints on the cosmological parameters are constantly improving as new observational data are collected. The shaded contours in Figure 5.12a illustrate the ranges of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that are consistent at different significance levels with a recent analysis of 580 Type Ia supernovae. The supernova-derived constraints *alone* show very strong evidence that $\Omega_{\Lambda,0} > 0$.

In Chapter 11 you will learn about another observable phenomenon called baryon acoustic oscillations (BAOs) that allows the spatial distribution of galaxies in the Universe to be used as a standard rod. The contours in Figure 5.12b show the constraints on $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that BAOs can provide. BAOs and Type Ia supernovae exclude different values of the density parameters; by combining these complementary constraints, cosmologists can completely rule out a $\Omega_{\Lambda,0} = 0$ universe.

In the next chapter you will learn how the properties of the cosmic microwave background can further restrict the range of plausible models for the real Universe.

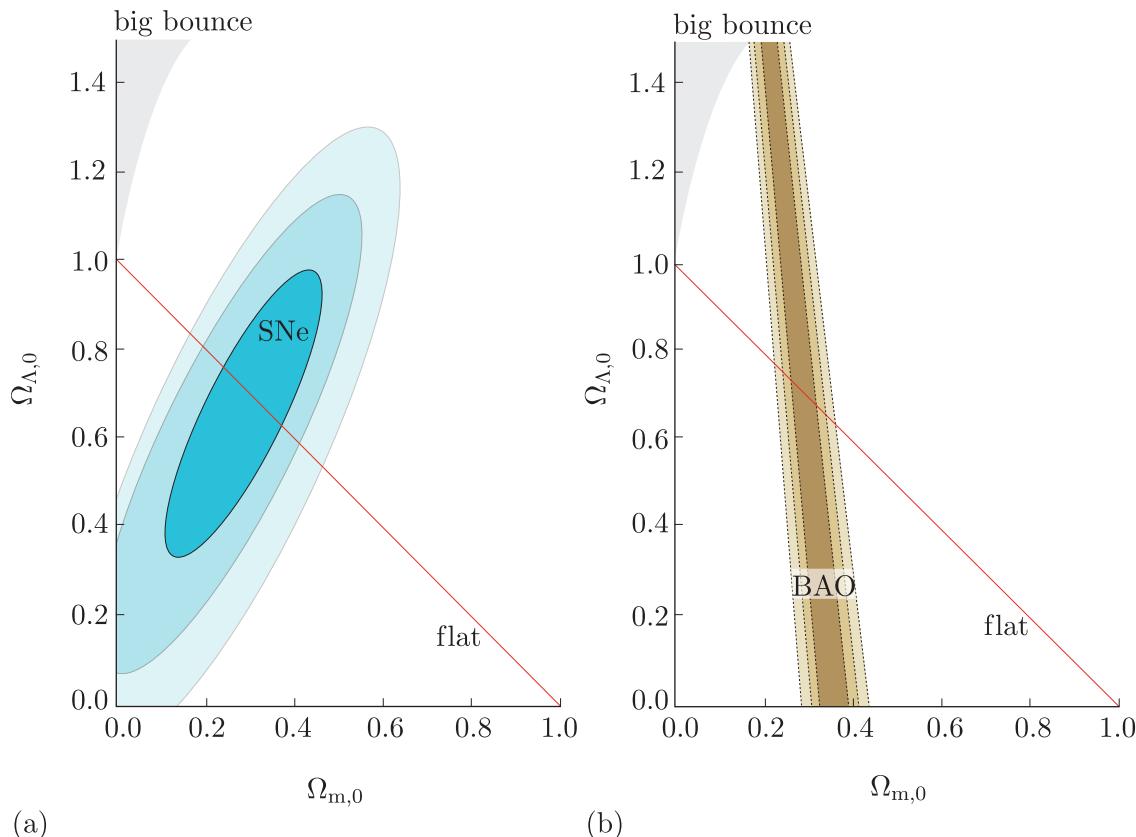


Figure 5.12 Constraints on $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ derived using observations of (a) Type Ia supernovae, and (b) baryon acoustic oscillations. The shaded contour levels indicate 1σ (darkest), 2σ and 3σ (lightest) uncertainty regions.

5.4 Summary of Chapter 5

- The **proper distance** $d_p(t)$ between two points measures the length of a spatial geodesic connecting them at a particular cosmic time, t :

$$d_p(t) = a(t) \int_0^R \frac{dr}{\sqrt{1 - kr^2}} \quad (\text{Eqn 5.1})$$

Assuming $a(t_0) = 1$, the *present* proper distance between two points is therefore equal to the co-moving distance d_c between them:

$$d_p(t_0) = a(t_0)d_c = d_c \quad (\text{Eqn 5.3})$$

- The **luminosity distance** d_L relates the observed flux F of a distant object to its intrinsic luminosity L :

$$d_L = \sqrt{\frac{L}{4\pi F}} \quad (\text{Eqn 5.11})$$

If $a(t)$ is *known* then the measured luminosity distance to a distant object can be used to calculate its intrinsic luminosity.

- The **angular diameter distance** d_A relates the observed angular size θ of a distant object to its physical size l .

$$d_A = \frac{l}{\sin \theta} \approx \frac{l}{\theta} \quad (\text{Eqn 5.13})$$

If $a(t)$ is *known* then the measured angular diameter distance to a distant object can be used to calculate its physical size.

- In a spatially flat universe, the present proper distance to a distant object, its luminosity distance and its angular diameter distance can be related in terms of its redshift z :

$$d_p(t_0) = \frac{d_L}{1+z} = (1+z)d_A \quad (5.38)$$

- Estimates of the proper distance to objects that have a wide range of redshifts can be used to measure $a(t)$, and thereby constrain cosmological parameters like $\Omega_{m,0}$, $\Omega_{\Lambda,0}$ and H_0 .
- Stellar parallax** can be used to directly measure the proper distance to objects within the Milky Way and its satellite galaxies.
- Objects with *known* intrinsic luminosity are called **standard candles**. Measuring the *observed* brightness of standard candles allows their luminosity distances to be determined. These luminosity distances can be used to calculate the corresponding proper distances if the standard candles' redshifts are known.
- Classical Cepheid** (CC) stars can be used as standard candles because their mean **absolute magnitude** $\langle M \rangle$ can be inferred from measurements of their pulsation periods P using a **period–luminosity relation**:

$$\langle M \rangle = A [\log_{10}(P) - 1] - B \quad (\text{Eqn 5.24})$$

To calibrate the relation, the coefficients A and B must be empirically determined by observing nearby CCs whose distances can be directly measured using stellar parallax. Periods have been measured for CCs at distances up to 33 Mpc, where the **peculiar velocities** of their host galaxies are comparable to the recession velocities that result from the expansion of the Universe.

- All **Type Ia supernovae** explode with similar, but not identical, peak luminosity. Even though their luminosities are not identical, they can still be used as standard candles because the *shapes* of their **light curves** can be used to reliably ‘correct’ their peak apparent magnitudes, and make them reflect the apparent peak magnitude of an idealised ‘standard’ Type Ia supernova.
- Type Ia supernovae have been observed and used as standard candles out to redshifts $z \approx 2$, where their velocities are dominated by the Hubble flow.
- Cosmologists use a technique known as the **cosmological distance ladder**, which is ‘anchored’ using measured stellar parallaxes, to calibrate the CC period–luminosity relation and the relation between a Type Ia supernova’s light curve shape and its intrinsic luminosity.
- The corrected peak *apparent* magnitudes of Type Ia supernova populations at low and high redshift can be used in combination to directly measure H_0 if the *absolute* magnitude of the idealised ‘standard’ supernova is known.
- In addition, the corrected peak *apparent* magnitudes of such supernova populations can be used in combination to measure $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$, even if H_0 is not known.

Solutions to exercises

Solution to Exercise 1.1

As set out in Equation 1.4, the surface brightness we observe for an object of a particular angular size depends on its flux. So we must first calculate the Sun's flux from its luminosity using Equation 1.3:

$$\begin{aligned} F_{\odot} &= L_{\odot}/4\pi D^2 \\ &= (3.8 \times 10^{26} \text{ W})/(4\pi(1.5 \times 10^{11} \text{ m})^2) \end{aligned}$$

which works out to be approximately 1340 W m^{-2} . We can now use Equation 1.4 to determine Σ_{\odot} :

$$\begin{aligned} \Sigma_{\odot} &= \frac{F_{\odot}}{\theta_{\odot}^2} \\ &= (1340 \text{ W m}^{-2})/(0.5 \text{ deg})^2 \\ &= 5400 \text{ W m}^{-2} \text{ deg}^{-2} \end{aligned}$$

Comparing this surface brightness with the value provided for the night sky, $\sim 2 \times 10^{-13} \text{ W m}^{-2} \text{ deg}^{-2}$, we conclude that the Sun's surface brightness is $\sim 3 \times 10^{16}$ times brighter than the night sky.

Solution to Exercise 1.2

The mean free path for a photon in the halo of the Milky Way at the present time is given by

$$\begin{aligned} \lambda &= \frac{1}{(100 \text{ m}^{-3})(6.652 \times 10^{-29} \text{ m}^2)} \\ &= 1.50 \times 10^{26} \text{ m} \end{aligned}$$

The mean free path for a photon at the given epoch in the early Universe is

$$\begin{aligned} \lambda &= \frac{1}{(5.0 \times 10^9 \text{ m}^{-3})(6.652 \times 10^{-29} \text{ m}^2)} \\ &= 3.01 \times 10^{18} \text{ m} \end{aligned}$$

Converting both values into units of kpc gives $4.9 \times 10^6 \text{ kpc}$ and 0.097 kpc , respectively.

Hence the mean free path in the halo of the Milky Way is much larger than the size of a typical galaxy, whereas in early Universe conditions the mean free path was much smaller than a typical galaxy.

Solution to Exercise 1.3

We start from the simplified form of the ionisation fraction equation:

$$\frac{1-X}{X} = n_p C(T)$$

which rearranges to:

$$1 = n_p C(T) X + X = X(n_p C(T) + 1)$$

and so

$$X = \frac{1}{n_p C(T) + 1}$$

We can now substitute in the values provided for each case. First, evaluating $C(T)$:

$$C(T) = \left(\frac{m_e k_B T}{2\pi \hbar^2} \right)^{-3/2} \exp \left(\frac{Q}{k_B T} \right)$$

and taking care to use $\hbar = h/(2\pi)$, gives values of $4.8 \times 10^{-31} \text{ m}^3$, $2.2 \times 10^{-10} \text{ m}^3$ and $1.8 \times 10^{-4} \text{ m}^3$ for scenarios (a), (b) and (c), respectively.

Using our values of $C(T)$ and the given values of n_p we find that $X = 1.0$, 0.48 and 1.1×10^{-6} (i.e. close to zero), respectively, for situations (a), (b) and (c). Note that (as will be the case throughout the module) your calculations may differ very slightly depending on the precision of constants and any intermediate rounding.

Solution to Exercise 1.4

We can use the provided proportionality to compare ratios of quantities at different times:

$$\frac{T_{\text{CMB}}}{T_{\text{present}}} = \frac{1+z_{\text{CMB}}}{1+z_{\text{present}}}$$

We know the temperature of the CMB at the time of emission, $T_{\text{CMB}} \approx 3000 \text{ K}$, and its temperature as observed at the Earth in the present day, $T_{\text{present}} = 2.7 \text{ K}$. The current redshift is zero. We can therefore rearrange the equation to obtain

$$z_{\text{CMB}} = \frac{T_{\text{CMB}}}{T_{\text{present}}} - 1$$

and substituting in the provided values gives $z_{\text{CMB}} = 1110$.

Solution to Exercise 2.1

(a) If $V = 20 \text{ km h}^{-1} = 5.6 \text{ m s}^{-1}$ then, using Equation 2.5, $\gamma = 1.0$.

(b) For $V = 0.9c$ you should have determined that $\gamma = 2.3$.

If you tried the optional Python task then you should have produced a plot similar to Figure S1.

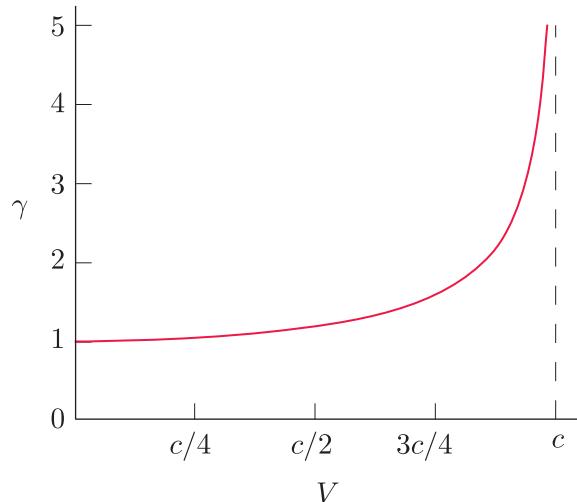


Figure S1 The dependence of the Lorentz factor, γ , on speed V .

Solution to Exercise 2.2

Equations 2.6 and 2.7 state

$$\begin{aligned}\Delta x' &= \gamma(\Delta x - V\Delta t) \\ \Delta t' &= \gamma(\Delta t - V\Delta x/c^2)\end{aligned}$$

To express these equations in terms of coordinates in the S' frame, first rearrange the second equation for Δt :

$$\Delta t = \frac{\Delta t'}{\gamma} + \frac{V\Delta x}{c^2}$$

This expression includes Δx , which is not a measurement from S' , so we need to rearrange the other interval transform so that we can eliminate Δx :

$$\Delta x = \frac{\Delta x'}{\gamma} + V\Delta t$$

Substituting this for Δx in the previous expression:

$$\Delta t = \frac{\Delta t'}{\gamma} + \frac{V(\Delta x'/\gamma + V\Delta t)}{c^2}$$

Collecting the Δt terms together gives:

$$\Delta t(1 - V^2/c^2) = \frac{1}{\gamma}(\Delta t' + V\Delta x'/c^2)$$

Recognising that $1 - V^2/c^2 = 1/\gamma^2$ and rearranging gives:

$$\Delta t = \gamma(\Delta t' + V\Delta x'/c^2)$$

which is an expression for Δt only in terms of S' coordinates.

A similar process leads to an expression for Δx in terms of only $\Delta x'$ and $\Delta t'$: $\Delta x = \gamma(\Delta x' + V\Delta t')$

Compared to the starting expressions for $\Delta t'$ and $\Delta x'$, you will see that only the sign of V has changed. This makes sense, because there is no reason we could not have labelled the two frames the other way round initially, in which case the direction of the velocity would have changed sign.

Solution to Exercise 2.3

Start with an expression for ds' in the situation where the y - and z -separations can be neglected:

$$ds'^2 = c^2 dt'^2 - dx'^2$$

The aim is now to show this is the same as ds^2 .

We can take dt' and dx' to be equivalent to the space and time intervals referred to as $\Delta t'$ and $\Delta x'$ in Equations 2.6 and 2.7, and so substitute in:

$$dx' = \gamma(dx - V dt)$$

and

$$c dt' = \gamma(c dt - V dx/c)$$

where we have multiplied the differential form of Equation 2.7 by c , so that

$$ds'^2 = [\gamma(c dt - V dx/c)]^2 - [\gamma(dx - V dt)]^2$$

which simplifies to

$$ds'^2 = \gamma^2 c^2 dt^2 (1 - V^2/c^2) - \gamma^2 dx^2 (1 - V^2/c^2)$$

Recalling the definition of $\gamma = 1/\sqrt{1 - V^2/c^2}$, this becomes

$$ds'^2 = c^2 dt^2 - dx^2 = ds^2$$

as required.

Solution to Exercise 2.4

The distance along a path on the surface of a sphere can be calculated by integrating the line element dl along a parametrisation of the path. For a line of constant latitude, this is very similar to the calculation in Example 2.6 because lines of latitude have a constant angle θ if the coordinate system is defined as in Figure 2.11, so that the angle ϕ corresponds to longitude, and θ corresponds to 90° minus the latitude value.

The distance, L , is therefore given by

$$L = \int_{\phi_1}^{\phi_2} R \sin \theta \, d\phi$$

where ϕ_1 and ϕ_2 are the start and end values of longitude, and $R = 6370$ km is the (average) radius of the Earth.

We need to work with angles in radians (rad), and so $\theta = 90^\circ - 51^\circ = 39^\circ$, which corresponds to 0.681 rad. The actual longitude angles were not provided, but the location of the zero value doesn't affect the calculation, so we can define $\phi_1 = 0^\circ$ and $\phi_2 = 118^\circ = 2.06$ rad.

Solving the integral as in Example 2.6 gives

$$\begin{aligned} L &= R \sin \theta [\phi]_{\phi_1}^{\phi_2} = R \sin \theta \phi_2 \\ &= (6370 \text{ km})(\sin(0.681))(2.06) = 8260 \text{ km} \end{aligned}$$

So the distance along the route at fixed latitude is 8260 km, which is around 13 per cent longer than the great-circle distance of ~ 7300 km.

Solution to Exercise 2.5

(a) We can use Equation 2.17 to calculate the radius corresponding to the provided value of k_x :

$$R = \frac{1}{k_x} = \frac{1}{0.2 \text{ cm}^{-1}} = 5 \text{ cm}$$

Because the curvature, and hence R , are the same at all locations, the curve must be a circle of radius 5 cm.

(b) A straight line must have constant curvature. But however large a circle we draw tangential to the line, an even larger circle will approximate the straight line better. The curvature of a straight line must therefore be smaller than $1/R$ for all R ,

i.e. $k_x = 0$. Mathematically,

$$k_x = \lim_{R \rightarrow \infty} \frac{1}{R} = 0$$

Solution to Exercise 3.1

The second half of the expression, ‘matter tells space how to curve’, is a rough description of Einstein's field equations: the energy–momentum tensor is the source of curvature and determines the components of the curvature tensor.

The first half of the expression, ‘space tells matter how to move’, refers to what a metric theory of gravity means: the metric (‘space’ in the colloquial expression) determines the geodesics along which, in the absence of external forces, test particles (‘matter’) must move.

Solution to Exercise 3.2

Using Equation 3.3, we find:

Table S1 Schwarzschild and physical radii values for three objects.

Object	R_S/m	R/m
Earth	8.9×10^{-3}	6.4×10^6
Sun	3.0×10^3	7.0×10^8
neutron star	5.9×10^3	1.5×10^4

The table therefore shows that the Schwarzschild radii of both the Earth and the Sun are much smaller than their actual size. For a neutron star – the densest type of star known – the Schwarzschild radius is still less than the star's actual radius, but only by a small factor.

Solution to Exercise 3.3

(a) A black hole of mass $3 M_\odot$ has a Schwarzschild radius of 9 km. Substituting this into Equation 3.6 gives a time to reach the singularity of 2×10^{-5} s.

(b) A 10^9 -solar-mass black hole has a Schwarzschild radius of $\sim 3 \times 10^{12}$ m, which results in a travel time of ~ 7000 s.

Solution to Exercise 3.4

Taking the scale factor at the present time, $a(t_0)$, to be 1, and $H(t) = H_0$, Equation 3.11 rearranges to

$$\frac{da}{dt} = H_0$$

and so

$$\int_{a(t)}^{a(t_0)} da = H_0 \int_t^{t_0} dt$$

so that

$$a(t_0) - a(t) = 1 - a(t) = H_0(t_0 - t)$$

So for part (a) we know that $a(t) = 1 - H_0(1000 \text{ y})$. To evaluate this, the units of H_0 and t need to be converted to match each other.

H_0 can be expressed in units of s^{-1} as follows:

$$H_0 = \frac{68\,000 \text{ m s}^{-1} \text{ Mpc}^{-1}}{3.086 \times 10^{22} \text{ m Mpc}^{-1}} = 2.2 \times 10^{-18} \text{ s}^{-1}$$

Likewise, 1000 years is $3.2 \times 10^{10} \text{ s}$, while 100 million years is $3.2 \times 10^{15} \text{ s}$.

Hence for part (a), $a(t) = 1 - 7.0 \times 10^{-8}$, which is very close to 1, and for part (b), $a(t) = 1 - 0.0070 = 0.993 \approx 0.99$ to the accuracy of this calculation.

You can therefore conclude that, with the assumption that H_0 is constant, the scale factor changed by a tiny factor over the past millennium, and by a little less than 1% ($\sim 0.7\%$) over the timescale of 100 million years.

Solution to Exercise 3.5

Using the relationship between frequency and Δt , Equation 3.14 becomes

$$\frac{1}{a(t_{\text{em}})\nu_{\text{em}}} = \frac{1}{a(t_{\text{obs}})\nu_{\text{obs}}}$$

which rearranges to give a relation for the ratio between scale factors at the time light was emitted and observed:

$$\frac{a(t_{\text{obs}})}{a(t_{\text{em}})} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}}$$

Recognising that the right-hand ratio is related to redshift via Equation 3.13, we can obtain the

required relationship between a and z :

$$\frac{a(t_{\text{obs}})}{a(t_{\text{em}})} = 1 + z \quad (\text{S1})$$

Solution to Exercise 3.6

The ratios of the scale factor at the time the light from each galaxy was emitted to its value at the present time are calculated via Equation 3.15 to be $a_{\text{em}}/a_0 = 1/(1+10.1) = 0.09$ and $a_{\text{em}}/a_0 = 1/(1+3.6) = 0.22$.

Therefore, the distance between any two locations was around 9% of its current distance at $z = 10.1$ and around 22% of its current distance at $z = 3.6$.

Solution to Exercise 4.1

For a group of n non-relativistic particles, each with energy E , contained within an arbitrary volume V of the homogeneous fluid, we can write the energy density as

$$\epsilon = \frac{nE}{V} \quad (\text{S2})$$

Now, if $v \ll c$ then $p \approx m_0 v$ (where v is the speed of the particles), and we can write the particles' energy as

$$E = \sqrt{p^2 c^2 + m_0^2 c^4} \approx m_0 c^2 \sqrt{1 + \frac{v^2}{c^2}}$$

Taking the first two terms in the Taylor expansion for $\sqrt{1+x}$ with $x = v^2/c^2$, and remembering that $v \ll c$, yields

$$E \approx m_0 c^2 + \frac{1}{2} m_0 v^2 \approx m_0 c^2$$

Substituting for E in Equation S2 we can rewrite the energy density of our particles as

$$\epsilon = \frac{nE}{V} \approx \frac{nm_0 c^2}{V} \approx \rho c^2$$

Solution to Exercise 4.2

(a) We are told that the WHIM behaves like an ideal gas, so we can use the ideal gas law to compute an approximate value for w . Comparing the form of Equation 4.13 with that of Equation 4.11, substituting μ for the proton mass m_p and using the proton temperature T_p , we can

compute w using

$$\begin{aligned} w &= \frac{k_B T}{\mu c^2} = \frac{k_B T_p}{m_p c^2} \\ &= \frac{1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^7 \text{ K}}{1.673 \times 10^{-27} \text{ kg} \times (2.998 \times 10^8 \text{ m s}^{-1})^2} \\ &= 9.184 \times 10^{-7} \end{aligned}$$

So $w < 10^{-6}$, which agrees with the assertion made in the question.

(b) We have shown that $w < 10^{-6}$ for protons in the WHIM, but we ultimately want to determine whether their velocities are non-relativistic.

If we assume that the gas *is* non-relativistic, then the average kinetic energy of a particle is $(1/2)\langle \mu v_p^2 \rangle = (1/2)\mu \langle v_p^2 \rangle$. We can then write:

$$3k_B T_p = \mu \langle v_p^2 \rangle = m_p \langle v_p^2 \rangle \quad (\text{S3})$$

where $\langle v_p^2 \rangle$ is the mean squared velocity of the protons in the WHIM. If we find that $\langle v^2 \rangle \ll c^2$ then our initial assumption will have been justified. Rearranging Equation S3 we find:

$$\begin{aligned} \langle v_p^2 \rangle &= \frac{3k_B T_p}{m_p} \\ &= \frac{3 \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^7 \text{ K}}{1.673 \times 10^{-27} \text{ kg}} \\ &= 2.476 \times 10^{11} \text{ J kg}^{-1} \\ &= 2.476 \times 10^{11} \text{ m}^2 \text{ s}^{-2} \\ &\approx 3 \times 10^{-6} c^2 \end{aligned}$$

(c) The previous step shows that $\langle v^2 \rangle$ is indeed much smaller than c^2 in this example. This result justifies the assumption that even the superheated protons in the WHIM are non-relativistic.

Solution to Exercise 4.3

When $k = 0$, the Friedmann equation becomes a separable first-order differential equation. To solve for $a(t)$, we start by rearranging Equation 4.16 to separate terms related to a and t :

$$a^{1/2} da = \left(\frac{8\pi G}{3} \rho_{m,0} \right)^{1/2} dt$$

The next step is to integrate both sides, and after collecting constants of integration we obtain

$$\frac{2}{3} a^{3/2} = \left(\frac{8\pi G}{3} \rho_{m,0} \right)^{1/2} t + \cancel{\text{constant}}^0$$

In case it is unfamiliar, the notation \cancel{X}^0 means ‘ X cancels to zero’. In this instance, we can set the integration constant to zero by assuming that the universe in question has zero size at the big bang when $t = 0$, and therefore that $a(0) = 0$. The expression can then be rearranged to obtain

$$a = (6\pi G \rho_{m,0})^{1/3} t^{2/3}$$

Finally, we use the boundary condition that $a(t_0) = 1$ to show that

$$a(t) = \left(\frac{t}{t_0} \right)^{2/3}$$

when $t_0 = (6\pi G \rho_{m,0})^{-1/2}$, as required.

Solution to Exercise 4.4

We can use exactly the same approach as was used for a matter-only universe. To solve for $a(t)$, we start by rearranging Equation 4.21 to separate terms related to a and t :

$$a da = \left(\frac{8\pi G}{3} \rho_{r,0} \right)^{1/2} dt$$

The next step is to integrate both sides, and after collecting constants of integration we obtain

$$\frac{a^2}{2} = \left(\frac{8\pi G}{3} \rho_{r,0} \right)^{1/2} t + \cancel{\text{constant}}^0$$

As in the matter-only case, we can set the integration constant equal to zero by assuming that the universe in question has zero size at the big bang when $t = 0$, and therefore that $a(0) = 0$. Then this expression can be rearranged to obtain

$$a = \left(\frac{32\pi G \rho_{r,0}}{3} \right)^{1/4} t^{1/2}$$

Finally, we use the boundary condition that $a(t_0) = 1$ to show that

$$a(t) = \left(\frac{t}{t_0} \right)^{1/2}$$

when $t_0 = (32\pi G \rho_{r,0}/3)^{-1/2}$.

Solution to Exercise 4.5

In general, the curvature parameter k can be positive, negative or zero. If a universe is static then $\dot{a} = 0$, and Equation 4.34 can be written as

$$\frac{8\pi G}{3}\rho + \frac{\Lambda c^2}{3} = \frac{kc^2}{a^2}$$

You are told to consider a universe that is not empty, so $\rho > 0$, and the question indicates that $\Lambda > 0$. Consequently, both terms on the left-hand side of this rewritten equation are non-zero and positive, so the right-hand side must also be positive. We also know that $a > 0$ except when $t = 0$ at the big bang, so the curvature k must be positive too.

Solution to Exercise 5.1

If we adopt the standard convention that $a(t_0) = 1$, then we can simplify Equation S1 from Chapter 3 and relate the redshift of photons that we detect today to the scale factor of the Universe at the time t , when they were emitted:

$$1+z = \frac{1}{a(t)} \quad (\text{S4})$$

The Friedmann equation is given in Equation 4.41 as:

$$\frac{H^2}{H_0^2} = \frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \Omega_{\Lambda,0} + \frac{\Omega_{k,0}}{a^2}$$

Substituting for a using Equation S4 then taking the square root and rearranging, we find

$$\begin{aligned} H &= H_0 [\Omega_{m,0}(1+z)^3 + \Omega_{r,0}(1+z)^4 \\ &\quad + \Omega_{\Lambda,0} + \Omega_{k,0}(1+z)^2]^{1/2} \\ &= H_0 E(z) \end{aligned} \quad (\text{S5})$$

Now, by differentiating Equation S4 and rearranging, we find:

$$da = -a(t)^2 dz$$

We also know that:

$$H = \frac{\dot{a}}{a} = \frac{1}{a(t)} \frac{da}{dt}$$

Rearranging this expression and substituting for H using Equation S5 gives:

$$dt = \frac{da}{Ha(t)} = -\frac{a(t)}{E(z)H_0} dz$$

Finally, substituting for dt in Equation 5.7 gives:

$$\begin{aligned} d_p(t_0) &= c \int_{t_{\text{em}}}^{t_0} \frac{dt}{a(t)} \\ &= -c \int_{z(t_{\text{em}})}^{z(t_0)} \frac{a(t)}{a(t)E(z)H_0} dz \\ &= -\frac{c}{H_0} \int_{z_{\text{em}}}^0 \frac{dz}{E(z)} \\ &= \frac{c}{H_0} \int_0^{z_{\text{em}}} \frac{dz}{E(z)} \end{aligned}$$

Solution to Exercise 5.2

Consider a particular object at redshift z . By rearranging and combining Equations 5.12 and 5.14, the current proper distance to the object is related to d_L and d_A by:

$$d_p(t_0) = (1+z) d_A = \frac{d_L}{1+z}$$

Now let the observed angular size and flux of the object be θ_{obs} and F_{obs} , respectively. Using the definitions of d_L and d_A , we obtain:

$$(1+z) \frac{l}{\theta_{\text{obs}}} = \frac{1}{1+z} \sqrt{\frac{L}{4\pi F_{\text{obs}}}} \quad (\text{S6})$$

We are interested in demonstrating proportionality, so let's approximate the observed appearance of our object as a disc with diameter θ_{obs} . In that case, the solid angle subtended by the disc is proportional to θ_{obs}^2 . Rearranging Equation S6, we find the required proportionality relationship:

$$\frac{F_{\text{obs}}}{\theta_{\text{obs}}^2} = \frac{L}{4\pi l^2} \frac{1}{(1+z)^4} \quad (\text{S7})$$

Solution to Exercise 5.3

We will assume that d_b is the mean Earth–Sun distance, i.e. $d_b \approx 1 \text{ AU}$. Rearranging Equation 5.19 to isolate θ_{ϖ} we find:

$$\theta_{\varpi} = \frac{d_b}{d_p} \approx \frac{1 \text{ AU}}{163\,000 \text{ ly}}$$

Performing the calculation and converting units appropriately we find that:

$$\begin{aligned} \theta_{\varpi} &\approx 0.97 \times 10^{-10} \text{ radians} \\ &\approx 0.02 \text{ milliarcseconds} \\ &\approx 2 \times 10^{-5} \text{ arcseconds} \end{aligned}$$

Solution to Exercise 5.4

The value of H_0 provided has *dimensions* of inverse time (i.e. s^{-1}), but its *unit* is expressed in terms of two different length scales: Mpc and km. To eliminate the unit's dependence on distance we simply divide by the number of km in 1 Mpc, which is approximately 3.086×10^{19} .

This gives

$$\begin{aligned} H_0 &\approx \frac{67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}}{3.086 \times 10^{19} \text{ km Mpc}^{-1}} \\ &= 2.194 \times 10^{-18} \text{ s}^{-1} \end{aligned}$$

So the Hubble time is approximately

$$t_H \approx \frac{1}{2.194 \times 10^{-18} \text{ s}^{-1}} = 4.558 \times 10^{17} \text{ s}$$

Now we just need to convert this to My. There are 31 557 600 seconds in 1 year, so

$$t_H = \frac{4.558 \times 10^{17}}{31 557 600 \times 10^6} = 14 443 \text{ My}$$

or approximately 14.5 billion years. This value is on a similar scale to the current best estimate of the Universe's true age, which is around 13.7 billion years.

Solution to Exercise 5.5

By rearranging Equation 5.25 to isolate the apparent magnitude we can write:

$$\begin{aligned} m &= M + 5 \log_{10}(d_L/\text{pc}) - 5 \\ &= M + 5 \log_{10}\left(\frac{d_L/\text{pc}}{10/\text{pc}}\right) \end{aligned}$$

The units of H_0 mean that it will be more convenient if we express d_L in Mpc and so our previous expression becomes:

$$\begin{aligned} m &= M + 5 \log_{10}\left(\frac{d_L/\text{Mpc}}{10^{-5}/\text{Mpc}}\right) \\ &= M + 5 \log_{10}(d_L/\text{Mpc}) + 25 \end{aligned}$$

Now we use an algebraic trick of multiplying d_L by H_0/H_0 and rearranging:

$$\begin{aligned} m &= M + 5 \log_{10}\left(\frac{H_0}{H_0} d_L/\text{Mpc}\right) + 25 \\ &= M - 5 \log_{10} H_0 + 5 \log_{10}(H_0 d_L/\text{Mpc}) + 25 \end{aligned}$$

Solution to Exercise 5.6

We want to find the value of z for which:

$$1.1 \times cz = cz \left(1 + \frac{1 - q_0}{2} z\right)$$

Rearranging to isolate z , we find:

$$z = \frac{0.2}{1 - q_0}$$

We are told to assume that the Universe is flat and that $\Omega_{r,0}$ is negligible so we can use Equation 4.46 to compute q_0 .

$$\begin{aligned} q_0 &= \frac{\Omega_{m,0}}{2} - \Omega_{\Lambda,0} \\ &= \frac{0.3097}{2} - 0.6888 \\ &= -0.5340 \end{aligned}$$

Using this value for q_0 in Equation , we find:

$$z = \frac{0.2}{1 + 0.5340} = 0.1304$$

so the low-redshift approximation is 10% too low when $z = 0.13$.

Solution to Exercise 5.7

We know that all galaxies contain CCs, and that roughly one Type Ia supernova occurs per decade in a spherical volume with proper radius $r_p = 20 \text{ Mpc}$. We also know that if $z \ll 1$, then $d_p(t_0) \approx cz/H_0$. Using this approximation we can calculate that CCs are detectable within a volume that has proper radius:

$$\begin{aligned} r_p(z = 0.01) &\approx \frac{zc}{H_0} \approx \frac{0.01 \times 2.998 \times 10^5 \text{ km s}^{-1}}{67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}} \\ &\approx 44.3 \text{ Mpc} \end{aligned}$$

Now our expected supernova rate per decade is just the ratio of two spherical volumes

$$\frac{(44.3 \text{ Mpc})^3}{(20 \text{ Mpc})^3} \approx 11$$

We would expect to find just over one Type Ia supernova every year in a galaxy that contains detectable CCs.

References and acknowledgements

References

- Benedict, G. F. *et al.* (2007) ‘*Hubble Space Telescope* fine guidance sensor parallaxes of galactic Cepheid variable stars: period–luminosity relations’, *The Astronomical Journal*, 133(4), pp. 1810–1827. Available at <https://doi.org/10.1086/511980>.
- Betoule, M. *et al.* (2014) ‘Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples’, *Astronomy & Astrophysics*, 568, A22. Available at <https://doi.org/10.1051/0004-6361/201423413>.
- Hamuy, M. *et al.* (1993) ‘The 1990 Calán/Tololo supernova search’, *The Astronomical Journal*, 106(6), pp. 2392–2407. Available at <https://doi.org/10.1086/116811>.
- Kolb, U. (2010) *Extreme environment astrophysics*. Cambridge: Cambridge University Press, in association with The Open University.
- Lambourne, R. (2010) *Relativity, gravitation and cosmology*. Cambridge: Cambridge University Press, in association with The Open University.
- Liddle, A. (1999) *An introduction to modern cosmology*. Chichester: Wiley.
- Misner, C., Thorne, K. and Wheeler, J. (1973) *Gravitation*. San Francisco: W. H. Freeman.
- Newman, J. A. *et al.* (1999) ‘A Cepheid distance to NGC 4603 in Centaurus’, *The Astrophysical Journal*, 523(2), pp. 506–520. Available at <https://doi.org/10.1086/307764>.
- Perlmutter, S. *et al.* (1997) ‘Measurements of the cosmological parameters Ω and Λ from the first seven supernovae at $z \geq 0.35$ ’, *The Astrophysical Journal*, 483(2), pp. 565–581. Available at <https://doi.org/10.1086/304265>.
- Perlmutter, S. *et al.* (1997) ‘Measurements of Ω and Λ from 42 high-redshift supernovae’, *The Astrophysical Journal*, 517(2), pp. 565–586. Available at <https://doi.org/10.1086/307221>.
- Ryden, B. (2017) *Introduction to cosmology*. 2nd edn. New York: Cambridge University Press.
- Serjeant, S. (2010) *Observational cosmology*. Cambridge: Cambridge University Press, in association with The Open University.
- van Velzen, S. *et al.* (2021) ‘Seventeen tidal disruption events from the first half of ZTF survey observations: entering a new era of population studies’, *The Astrophysical Journal*, 908(4). Available at <https://doi.org/10.3847/1538-4357/abc258>.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: Max-Planck Institute for Physics.

Chapter images: Figure 1.1: M. Collness, Mount Stromlo Observatory; Figure 1.2: Hubble, E. (1929) ‘A relation between distance and radial velocity among extra-galactic nebulae’, *Proceedings of the National Academy of Sciences of the United States of America*, 15(3), pp. 168-173; Figure 1.3: Betoule, M. *et al.* (2014) ‘Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples’, *Cosmology and Nongalactic Astrophysics*, Cornel University, <https://arxiv.org/abs/1401.4064>, licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, <https://creativecommons.org/licenses/by/4.0/>; Figure 1.5: MissMJ, Cush, https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, licensed under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, <https://creativecommons.org/licenses/by/3.0/deed.en>; Figure 1.6: NASA/WMAP Science Team; Figure 1.8: ESA – C. Carreau, https://www.esa.int/ESA_Multimedia/Images/2013/03/Planck_history_of_Universe, licensed under a Creative Commons Attribution-ShareAlike 3.0 IGO (CC BY-SA 3.0 IGO) licence, <https://creativecommons.org/licenses/by-sa/3.0/igo/>; Figure 1.9: European Space Agency and the Planck Scientific Collaboration; Figure 3.6a: courtesy Caltech/MIT/LIGO Laboratory; Figure 3.6b: Kramer, M. *et al.* (2021), ‘Strong-field gravity tests with the double pulsar’, *Physics Review X*, 11:041050; Figure 3.7a: S. Gillessen, *et al.* (2009) ‘Monitoring stellar orbits around the massive black hole in the galactic centre’, *The Astrophysical Journal*, 692:10751109, American Astronomical Society, 2009; Figure 3.7b: EHT Collaboration, <https://www.eso.org/public/images/eso1907a/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 3.7c: V. Heesen and LOFAR surveys team; Figure 3.8: courtesy Caltech/MIT/LIGO Laboratory; Figure 3.11a: NASA/CXC/U. Michigan/J. Miller *et al.*, illustration NASA/CXC/M. Weiss; Figure 3.11b: Velzen, S. N. *et al.* (2021) ‘Seventeen tidal disruption events from the first half of ZTF survey observations: entering a new era of population studies’, *The Astrophysical Journal*, 908(1), The American Astronomical Society; Figure 5.6: ESA-D. Ducros, 2013; Figure 5.7: NASA/ESA, The Hubble Key Project Team and The High-Z Supernova Search Team, <https://esahubble.org/images/opo9919i/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 5.8: copyright unknown; Figure 5.12: Suzuki, N. (2012) ‘The Hubble Space Telescope Cluster Supernova Survey. V. Improving the dark-energy constraints above $z > 1$ and building an early-type-hosted supernova sample’, © 2012, The American Astronomical Society, all rights reserved, printed in the U.S.A.

Software: ‘Python’ and the Python logos are trademarks or registered trademarks of the Python Software Foundation, used by The Open University with permission from the Foundation.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Book production contributors

Academic authors

Judith Croston (Chair), Hugh Dickinson, Iain McDonald and Sheona Urquhart.

The authors would like to thank Stephen Serjeant, Bonny Barkus, Kate Gibson and Mark Jones for useful feedback and discussions.

External assessor

Stephen Wilkins, University of Sussex.

Curriculum team

Jessica Bartlett and Shelah Survey.

Production team

Senior project manager

Jeni Aldridge.

Editors

Jonathan Martyn, Peter Twomey, Yon-Hee Kim and Lil Davies.
Mark Radford (Pepperhouse Editorial) and Jonathan Darch.

Graphics

Sha’ni Hirsch.

OU Library

James Salter.

Index

Note: **bold** page numbers indicate where terms are defined.

- 2dF Galaxy Redshift Survey 4
- absolute magnitude **134**, 140
- acceleration equation 94, **94**
- age of Universe 11
- angular diameter distance **126**
- BAOs *see* baryon acoustic oscillations
- baryogenesis **22**
- baryon **14**
- baryon acoustic oscillations (BAOs) 151
- baryonic matter **14**
- big bang model **20**, 21
- big bounce universe 114
- big crunch 90
- binary pulsar system 65
- binary star system 136
- binding energy 18
- black hole 67, **69**
 - evidence 69
 - merger 70
- black-body spectrum **24**
- blueshift 6
- boson **12**
- brightness of sky 11
- Calán/Tololo Supernova Survey 145
- Cartesian coordinates 46
- causality **36**, 40
- causally connected **40**
- CC *see* classical Cepheid
- Chandrasekhar limit **136**
- classical Cepheid (CC) **133**
- closed universe **91**
- cluster 4
- CMB *see* cosmic microwave background
- co-moving coordinates **78**
- coasting universe 113
- co-moving radius 92
- conservation of energy 93
- contents of Universe 12
- coordinate system **29**
- coordinate time 57
- coordinates
 - Cartesian 46
 - polar 43
 - spherical 46
- cosmic microwave background (CMB) 20, **23**, 76
- cosmic time **77**
- cosmic web 4
- cosmological constant **107**
- cosmological distance ladder **142**
- cosmological principle 4, **4**, 8, 76, 121
- cosmological redshift 121
- critical density **103**
- critical redshift 128
- curvature **45**, 54
 - negative 48
 - positive 48
- curvature parameter **48**, **79**, 90
- curvature scalar 63
- dark ages **22**
- dark energy **15**
- dark matter **15**
- dark star 69
- deceleration parameter **112**
- δ Cephei 134
- density 15
 - critical 103
- density parameter **105**
- differential notation 42
- distance modulus **135**
- Doppler shift 69, 121
- Earth 60, 61, 129, 131
- Einstein tensor **63**
- Einstein's field equations **62**
- electromagnetic force 12
- electromagnetic interaction **13**
- electron **13**
- electron number density 18
- energy 14
- energy density 91
- energy-momentum tensor **63**
- equation of state **95**
 - cosmological constant 109
 - matter 96

- radiation 99
- equation of state parameter **95**
- equivalence principle **60**
- escape speed **69**
- event **29**
- event horizon **69**, 71, 73
- Event Horizon Telescope 69
- expansion of Universe 7, 12
- experimental uncertainty 147
- fermion **12**, 13
- first law of thermodynamics 92
- first light **22**
- flare 74
- flat universe **91**
- fluid equation **92**, 93
- flux **9**
- flux density **9**
- free fall 59
- Friedmann equation **87**, 89, 105–107, 109
- fundamental observers **76**
- Gaia* 131
- galaxy evolution **22**
- Galilean transformations **32**
- gas 15
- general relativity 45, 54, 91
 - Einstein’s field equations 63
 - evidence 64
- general theory of relativity 12, 106
- geodesic **48**, 55
- gravitational lensing 65
- gravitational potential energy 93
- gravitational redshift 65, **75**
- gravitational waves **65**
- gravity
 - metric theory 62
- great circle **48**
- H–R diagram *see* Hertzsprung–Russell diagram
- hadron **14**
- Hertzsprung–Russell (H–R) diagram 133
- High-z Supernova Search Team 138
- homogeneous 4, 76, 88, 95
- horizon distance **119**, 120
- HST* *see* *Hubble Space Telescope*
- Hubble constant **6**, 7, 80
 - estimation 140
- misnomer 8
- Hubble diagram 6, 7, 150
- Hubble distance **139**
- Hubble flow **76**
- Hubble parameter 8, **81**, 103
- Hubble Space Telescope (HST)* 134
- Hubble time **139**
- Hubble–Lemaître law **6**, 121, 141
- hypersurface 77
- ideal gas law **16**, 95, 96
- inertial frame **28**
- inflation **22**
- instability strip **133**
- interval **33**
- invariance
 - spatial 90
- ionisation **18**
- ionisation fraction 18, 20
- isotropic 4, 76, 88, 95
- Kerr metric **75**
- Large Magellanic Cloud (LMC) 131, 134, 138
- Laser Interferometer Gravitational-wave Observatory (LIGO) 65
- last scattering **22**
- length contraction 35
- lepton **13**
- light 17
- light cone **40**, 55
- light curve **137**
- light-like geodesic 55
- light-year 5
- LIGO *see* Laser Interferometer Gravitational-wave Observatory
- line element **42**
- LMC *see* Large Magellanic Cloud
- local inertial frames **61**
- loitering universe 113
- lookback time **22**
- Lorentz factor **32**
- Lorentz transformations **32**, 33, 37
- luminosity **9**
- luminosity distance **122**
- LVT151012 70
- magnitude system 133
- manifold **45**

- mass density 15
 matter 14, 96
 matter–antimatter annihilation 102
 matter fluid 96
 Maxwell–Boltzmann distribution **15**
 mean free path **17**
 measured redshift 121
 measurement uncertainty 147
 megaparsec (Mpc) 5
 metric **44**, 51, 55
 metric coefficient **44**
 metric tensor **44**, 63
 metric theory of gravity 62
 Michelson–Morley experiment 29
 Milky Way 69, 78
 - peculiar velocity 131
 Minkowski metric **44**, 54, 79
 Minkowski space 56
 Minkowski spacetime 54

 negative curvature **48**
 neutrino **13**, 99
 neutron **13**
 NGC 1262 78
 NGC 4526 136
 NGC 4603 135
 non-Euclidean geometry **45**
 notation 10
 nucleosynthesis **22**
 null geodesic **55**
 number density
 - electron 18
 - particle 15
 observer **29**
 Olbers’ paradox **11**
 opacity **17**
 open universe **91**
 overdensity 92

 parsec (pc) 5, 131
 particle horizon **119**
 particle number density 15
 particles form **22**
 Pauli exclusion principle **12**
 peculiar velocity **121**
 - Milky Way 131
 perfect fluid **95**

 perihelion 64
 period–luminosity relation **134**
 photon **12**, 16, 18, 99
 plasma **13**, 15
 polar coordinates 43
 positive curvature **48**
 positron **14**
 postulates of special relativity 29
 Pound–Rebka experiment **65**
 precession **64**
 pressure gradient 94
 proper distance **118**
 proper radius *see* particle horizon
 proper time **55**, 57, 71
 proton **13**
 pseudo-Riemannian 54

 quark **13**
 quasar 65, 130

 radiation 17, 99
 radiation constant 17
 radiation fluid 99
 recession 5
 recession velocity 106
 recombination **22**
 redshift **6**, 24, 25, 81, 83, 120
 - cosmological 121
 - critical 128
 - measured 121
 reference frame **28**, 30, 34, 56, 61
 - transformation 31
 relativistic 16
 relativity of simultaneity **42**
 Ricci curvature 63
 Riemann curvature tensor **53**
 Riemannian geometry 53
 Robertson–Walker metric **79**, 90

 Saha equation **18**, 19
 scale factor **78**, 80, 83, 98, 99, 118
 Schwarzschild metric **67**, 68
 Schwarzschild radius **67**, 68
 SCP *see* Supernova Cosmology Project
 SDSS *see* Sloan Digital Sky Survey
 Shapiro delay **65**
 simultaneity 42
 singularity **71**, 76

- Sloan Digital Sky Survey (SDSS) 7
SN 1987A 138
SN 1994D 136
SN UDS10Wil 138
SNLS *see* Supernova Legacy Survey
space-like geodesic 55
spacetime 8, 12, 36, 42, 45, 62, 66, 76, 94
spacetime diagram 36, 38–41
spacetime separation 43
spatial invariance 90
special relativity 27
 applications 36
 consequences 33
 postulates 29
 terminology 29
spectral shift 6
speed of light 29
spherical coordinates 46
standard candle 122
 Type Ia supernova 136
Standard Model 12
standard rod 126
standard ruler *see* standard rod
standard yardstick *see* standard rod
standardisable candle 137
stellar parallax 130, 131, 134
stress energy 94
strong interaction 13
structure formation 22
sub-relativistic 15
Sun 11, 64, 129, 131
 notation 10
supercluster 4
supernova 7
 standard candle 136
 standardisable candle 137
 Type Ia 136, 138, 140
Supernova Cosmology Project (SCP) 138, 145, 150
Supernova Legacy Survey (SNLS) 7
surface brightness 9, 11, 128

TDE *see* tidal disruption event
tensor 44
thermal equilibrium 15, 24
Thomson cross-section 18
Thomson scattering 17
tidal disruption event (TDE) 74
time dilation 35, 75
time-like geodesic 55, 56
Type Ia supernova 136, 138, 140
 number of 138
 standard candle 136
 standardisable candle 137

uncertainty
 measurement 147
universality of free fall 59
Universe
 age 11
 contents 12
 expansion 7, 12
 model 95
universe
 closed 91
 flat 91
 open 91

Virgo interferometer 65

warm–hot intergalactic medium 96
WD *see* white dwarf
weak interaction 13
Weyl’s postulate 76
WHIM *see* warm–hot intergalactic medium
white dwarf (WD) 136, 136
world line 39, 55, 56

S385 Errata: Cosmology Part 1

Chapter 2

- p56. In the solution to Example 2.8 the penultimate equation should be:

$$d\tau^2 = dt^2 - [(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2]/c^2$$

- p56. In the final sentence of the solution to Example 2.8, the reference to dt should instead refer to dt' since it is in the primed frame that the particle is at rest.

Chapter 4

- p105. The reference to the *sign* of $\Omega(t)$ in first sentence of the paragraph below Equation 4.29 should instead refer to the sign of $\Omega(t) - 1$. Since densities must always be positive and $\Omega(t)$ is a ratio of two densities, its sign must always be positive.

Chapter 5

- p138. The footnote on this page incorrectly states that SN 1987A was a type Ia supernova. In fact, SN 1987A was a type II supernova caused by the core collapse of a blue supergiant.
- p141. In the first sentence of Section 5.3.2, the symbol for the corrected apparent magnitude should be m_{corr} and not M_{corr} as stated.
- p143. In Figure 5.9, the redshift ranges given for the “Type Ia supernovae and classical Cepheids” and “Type Ia supernovae in the Hubble flow” categories are incorrect. Classical Cepheids are only detectable at redshifts $\lesssim 0.01$ while Type Ia supernovae can still be detected at redshifts above 1. Therefore, the correct range for “Type Ia supernovae and classical Cepheids” is $z \lesssim 0.01$ and the correct range for “Type Ia supernovae in the Hubble flow” is $0.01 \lesssim z \lesssim 1$.
- p147. In Example 5.6 there is a mistake in the derivation of the expression for $\sigma_{\mathcal{M}_i}^2$. The correct derivation and its implications are laid out below.

Starting with

$$\mathcal{M} = m - 5 \log_{10}(cz)$$

the partial derivative of \mathcal{M} with respect to z is:

$$\frac{\partial \mathcal{M}}{\partial z} = -5 \log_{10}(e) \frac{c}{cz} = \frac{-5}{z \ln(10)}$$

so

$$\left(\frac{\partial \mathcal{M}}{\partial z} \right)^2 = 25 \left[\frac{1}{z \ln(10)} \right]^2$$

Using this result in conjunction with Equation 5.35 yields:

$$\sigma_{\mathcal{M}_i}^2 = \sigma_{m_{\text{corr},i}}^2 + 25 \left[\frac{\sigma_{z_i}}{z_i \ln(10)} \right]^2$$

This changes the values in Table 5.6 to

Supernova	$\sigma_{\mathcal{M}_i}$	$\sigma_{\mathcal{M}_i}^2$
1990O	0.246894	0.060957
1992ae	0.208214	0.043353
1992aq	0.233985	0.054749
1992bo	0.319865	0.102313
1992bp	0.188208	0.035422

and so $\sigma_{\mathcal{M}}^2 = 0.059$

Using this result, table 5.7 becomes

Supernova	$\sigma_{\mathcal{D}}$
1995aw	0.065741
1995ax	0.108857
1997L	0.126141
1997S	0.126273
1997aj	0.097758

Overall, these corrections increase the size of the error bars in Figure 5.10 but the conclusion remain the same.

S385

Cosmology and the distant Universe

Cosmology Part 2



This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2023. Second edition 2024.

Copyright © 2023, 2024 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd, 1 St. Katharine's Way, London, E1W 1UN (website www.cla.co.uk).

Open University materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using L^AT_EX.

Printed and bound in the United Kingdom by Halstan Printing Group, Amersham

ISBN 978 1 4730 3988 9

2.1

Contents

Introduction	1
Chapter 6 The cosmic microwave background	3
6.1 The surface of last scattering	3
6.2 Observing the sound of the Universe	6
6.2.1 CMB observatories and telescopes	6
6.2.2 The noisy microwave sky	8
6.3 CMB temperature fluctuations	15
6.3.1 Describing individual temperature fluctuations	16
6.3.2 Statistical descriptions of the CMB	22
6.3.3 The CMB angular power spectrum	23
6.4 Summary of Chapter 6	26
Chapter 7 Cosmology with the cosmic microwave background	29
7.1 Explaining the CMB angular power spectrum	30
7.1.1 Before the epoch of last scattering	31
7.1.2 Acoustic oscillations	33
7.1.3 Radiation driving	38
7.1.4 The origin of the temperature fluctuations	39
7.2 Cosmological parameter measurements	41
7.2.1 The acoustic scale	41
7.2.2 Measuring Ω_k	44
7.2.3 Measuring $\Omega_{b,0}$	46
7.2.4 Measuring $\Omega_{m,0}$	49
7.2.5 Measuring $\Omega_{\Lambda,0}$	50
7.2.6 Cosmological parameter degeneracies and combined constraints	52
7.3 Summary of Chapter 7	55
Chapter 8 Physics of the early Universe	57
8.1 Particle interactions in the early Universe	57
8.1.1 Basics of early-Universe interaction physics	58
8.1.2 Key interaction processes	62

8.2 A timeline towards the formation of atomic nuclei	67
8.2.1 The very early Universe and inflation	68
8.2.2 Neutrino decoupling and the baryon-to-photon ratio	70
8.2.3 The neutron-to-proton ratio	72
8.3 Big bang nucleosynthesis	74
8.3.1 Deuterium formation	74
8.3.2 Helium formation	77
8.3.3 Formation of heavier nuclei	79
8.4 Summary of Chapter 8	80
 Chapter 9 Early Universe physics meets observations	 83
9.1 From nuclei to atoms	83
9.1.1 Recombination	83
9.1.2 Decoupling of matter and radiation	86
9.2 Measuring primordial abundances	88
9.2.1 Abundances and metallicity	89
9.2.2 Deuterium abundances	91
9.2.3 Helium abundances	94
9.2.4 Lithium abundance	97
9.2.5 The cosmological significance of Ω_b	98
9.3 Measuring non-baryonic matter	99
9.3.1 Weighing galaxy clusters	99
9.3.2 Galaxy rotation curves	104
9.3.3 Dark matter, the CMB and structure formation	106
9.4 Summary of Chapter 9	107
 Chapter 10 Structure formation and the cosmic web	 109
10.1 Growth of density perturbations	109
10.1.1 Hubble flow	109
10.1.2 Collapse or expansion?	111
10.1.3 Limitations of the Jeans criterion	112
10.2 Collapse of dark-matter halos	113
10.2.1 Collapse of a spherical overdensity	113
10.2.2 Incorporating expansion of the Universe	115
10.2.3 Collapsing dark matter and baryons	118
10.2.4 Virialisation	120
10.3 Collapse in three dimensions	121
10.3.1 Collapse of an ellipsoid	121

10.3.2 Numerical simulations	122
10.3.3 Testing simulation predictions	125
10.4 Summary of Chapter 10	126
Chapter 11 Formation of stars and galaxies	129
11.1 From baryonic gas to stars	130
11.1.1 Baryons in dark-matter halos	130
11.1.2 Forming structure on galaxy scales	132
11.1.3 Gas cooling	132
11.1.4 Reionisation and Strömgren spheres	136
11.2 Key processes in galaxy evolution	139
11.2.1 Galaxy assembly	140
11.2.2 The cosmic cycle of matter	141
11.2.3 Feedback from stellar remnants and supermassive black holes	145
11.2.4 Chemical evolution of galaxies	147
11.3 Galaxies as cosmological probes	149
11.3.1 Galaxy number counts	149
11.3.2 Galaxy luminosity and stellar mass functions	150
11.3.3 Clustering and the two-point correlation function	154
11.3.4 Baryon acoustic oscillations	155
11.4 Summary of Chapter 11	158
Solutions to exercises	161
References and acknowledgements	169
Index	173

Introduction

In this book you will continue your study of cosmology, learning more about the observational evidence used to test cosmological models, and exploring the timeline of the evolution of the Universe, from immediately after the big bang to the formation of stars and galaxies.

There are six chapters in *Cosmology Part 2*, as described below.

- Chapter 6 provides an in-depth introduction to the cosmic microwave background radiation (CMB) and its observational properties.
- Chapter 7 explores the underlying physics of how the CMB was produced, and considers how its observed properties are used to measure cosmological density parameters.
- Chapter 8 examines particle interactions in the early Universe, and the evolution of matter towards forming the first nuclei.
- Chapter 9 discusses the creation of the first atoms, considers how observations of cosmic elemental abundances are used to test cosmological models, and describes the evidence for dark matter.
- Chapter 10 examines the formation, via gravitational collapse, of a cosmic web of large-scale structures in the Universe, which eventually evolved to form stars and galaxies.
- Chapter 11 completes the story by providing an overview of the formation of stars and galaxies within the cosmic web.

As with Part 1, the exercises in each chapter are an important element of your learning, with full solutions provided at the end of the book. Some exercises and examples involve using the Python programming language to perform calculations. Solutions to these problems are available from the module website in the form of Jupyter Notebooks.

The table of physical constants is repeated at the end of this book for use in your calculations, and definitions for terms highlighted in **bold** may be found in the module glossary. Where we think it could be helpful we have also included references back to equations or figures in the first *Cosmology* book, which you may find useful to revisit to remind yourself of the relevant underlying concepts.

Throughout the text, coloured boxes are again used to highlight particular types of information. Orange boxes highlight the most important equations and other key information. Turquoise boxes indicate additional information, such as reminders of concepts that you may have met in previous study, or ideas that are partly beyond the scope of the module but provide additional context. Blue boxes indicate where further, optional resources are available on the module website.

As with *Cosmology Part 1*, the material in this book was heavily influenced by earlier Open University module material, and other excellent textbooks. Please refer to the Part 1 ‘Introduction’ if you would like more information about key influences on this material.

Chapter 6 The cosmic microwave background

In Chapter 1 you learned about a radiation field called the cosmic microwave background (CMB) that permeates the entire Universe. In this chapter we will explore the observable characteristics of the CMB in much more detail and you will learn what they reveal about cosmic evolution and the cosmological parameters of the Universe.

In the previous chapter you learned that cosmologists can use observations of distant stars, galaxies and supernovae to measure the expansion rate of the Universe and thereby constrain the values of the cosmological parameters. In this chapter and the next you will learn that precise measurements of the CMB provide cosmologists with another, largely complementary way to measure the properties of the Universe.

At the end of the next chapter you will see that by combining the constraints derived from supernova observations with those derived from the CMB, cosmologists can derive very precise constraints for the cosmological parameters.

Objectives

Working through this chapter will enable you to:

- describe and understand the physical processes that produced the cosmic microwave background (CMB)
- explain the detailed observational properties of the CMB
- compare the capabilities of different telescopes that can be used to observe the CMB from the Earth and from space
- describe the challenges associated with observing the CMB
- summarise the ways that cosmologists mathematically describe and statistically model the spatial and spectral properties of the CMB.

6.1 The surface of last scattering

The Universe was almost completely opaque to electromagnetic radiation for the first $\sim 370\,000$ years of its evolution. During this time, the temperature of the Universe was much higher than it is today and typical photon energies were much larger than the ionisation potential Q of hydrogen. Consequently, the baryonic content of the Universe was almost completely ionised. Photons could only propagate for short distances before encountering a free electron and undergoing Thomson scattering, and any hydrogen atoms that did *briefly* form were almost immediately reionised.

Chapter 6 The cosmic microwave background

This situation persisted until cosmological expansion and redshift reduced enough photons' energies below Q , which allowed widespread and rapid formation of long-lived hydrogen and helium atoms. As the number of free electrons fell, so did the Thomson scattering rate. Very soon, most photons were able to stream freely through the Universe and most have continued to do so ever since. It is *these* photons, released during the short interval when the Universe first became transparent, that have propagated without interaction ever since and make up the CMB.

Later chapters describe in more detail the physics that governed particle interactions and the co-evolution of matter and radiation during the Universe's early history. A key concept that will be needed for *this* chapter is that the photons making up the CMB effectively encode information about the state of the Universe when it was less than half a million years old and looked very different from the way it does today.

The speed at which the Universe became transparent means that all of the photons that make up the CMB have travelled a very similar distance since the last time they underwent Thomson scattering. For this reason the CMB, as seen by a *particular* observer, is often described as having been emitted from a spherical **surface of last scattering**. The time in the Universe's history when a *typical* CMB photon last underwent Thomson scattering is called the **epoch of last scattering**, which we can abbreviate as t_{ls} . We will also use the notation z_{ls} to represent the redshift measured today for photons that were emitted at t_{ls} . The numerical value of z_{ls} is approximately 1090.

Observers at different locations in the Universe detect CMB photons from *different* surfaces of last scattering. Figure 6.1 shows that all such surfaces have the same proper radius, which we denote using the symbol $d_{\text{p}}(t_0, z_{\text{ls}})$. This notation represents the proper distance measured at the present time, t_0 , travelled by photons with observed redshift z_{ls} . Note that the surface of last scattering is closer than the present-day horizon distance $d_{\text{hor}}(t_0)$, so we can observe the *effects* of processes that occurred before t_{ls} . However, any photons that were emitted at proper distances larger than $d_{\text{p}}(t_0, z_{\text{ls}})$ are very unlikely to reach a present-day observer without having interacted, because they would have been emitted when the Universe was almost completely opaque.

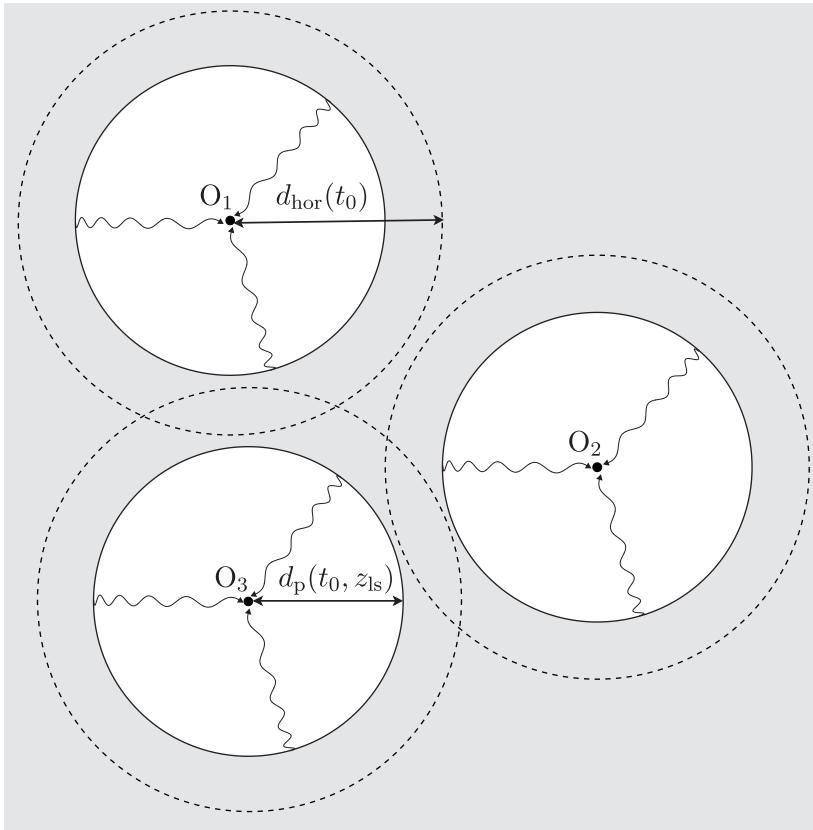


Figure 6.1 Three observers, O_1 , O_2 and O_3 , at three different, widely separated locations in the Universe at the present cosmic time t_0 .

In Figure 6.1, the solid circles represent three different surfaces of last scattering, each having a different observer at its centre. Each surface has the same proper radius $d_p(t_0, z_{ls})$, which is equal to the maximum proper distance that photons could have travelled since the epoch of last scattering. The dashed circles represent spherical surfaces with radii equal to the present horizon distance $d_{\text{hor}}(t_0)$.

In reality, the Universe's transition from opaque to transparent was not instantaneous. As we will see in the next chapter, different regions of the Universe were slightly denser or hotter than others so all would have become transparent at slightly different times. Accordingly, the last-scattering redshifts of CMB photons span a small range of values centred around z_{ls} , such that the 'surface' is actually more like a thin shell.

6.2 Observing the sound of the Universe

In Chapter 1 you saw that maps of the CMB temperature are remarkably uniform but not perfectly smooth. There are tiny differences between the CMB temperatures measured at different locations on the sky. As you continue to work through this chapter, you will learn that one of the primary goals when observing the CMB is to measure the amplitudes of these temperature fluctuations (also commonly referred to as **CMB anisotropies**) and the statistical distribution of their angular sizes. By analysing these measurements, cosmologists can recover a wealth of information about the cosmological parameters and the physical environment of the early Universe.

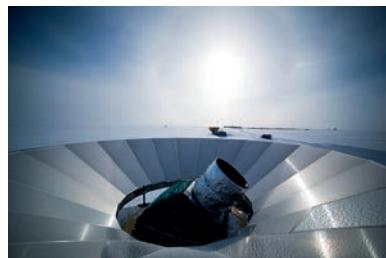
Fundamentally, the CMB temperature fluctuations reflect a complex pattern of overlapping sound waves with different wavelengths and amplitudes that propagated through the fluids that filled the early Universe. Cosmologists call these sound waves **acoustic oscillations**, and many of the mathematical and statistical techniques that we will use in this chapter have close analogues in the study of acoustics and other wave phenomena. The pattern of temperature fluctuations that we see in the CMB reflects the pattern of acoustic oscillations as it existed at the epoch of last scattering. In the next chapter you will learn much more about the physical processes that govern the detailed properties of the acoustic oscillations and, by extension, the CMB fluctuations.

Before they can derive any cosmological inferences from the CMB, cosmologists must first make detailed observations of its properties. In this section we will review the challenges that are associated with measuring the CMB's properties and the astonishing advances in telescope technology that have made such measurements possible.

6.2.1 CMB observatories and telescopes

As its name suggests, observing the CMB requires telescopes that can detect microwave radiation arriving from space. Observing the CMB from the ground is very challenging because water molecules in the Earth's atmosphere strongly absorb celestial microwaves and can even emit microwave radiation that resembles celestial signals.

To observe the microwave sky from the ground, astronomers must construct their telescopes at high-altitude sites that have very low atmospheric humidity. One location on Earth that fulfils these criteria is the Antarctic Plateau: this region has an average elevation of $\sim 3000\text{ m}$ and is kept very dry by the extreme cold, which prevents water vapour from forming. Since 1984, several microwave telescopes have been constructed very close to the South Pole; the three currently in operation are shown in Figure 6.2 – BICEP3, the South Pole Telescope and BICEP Array.



(a)



(b)

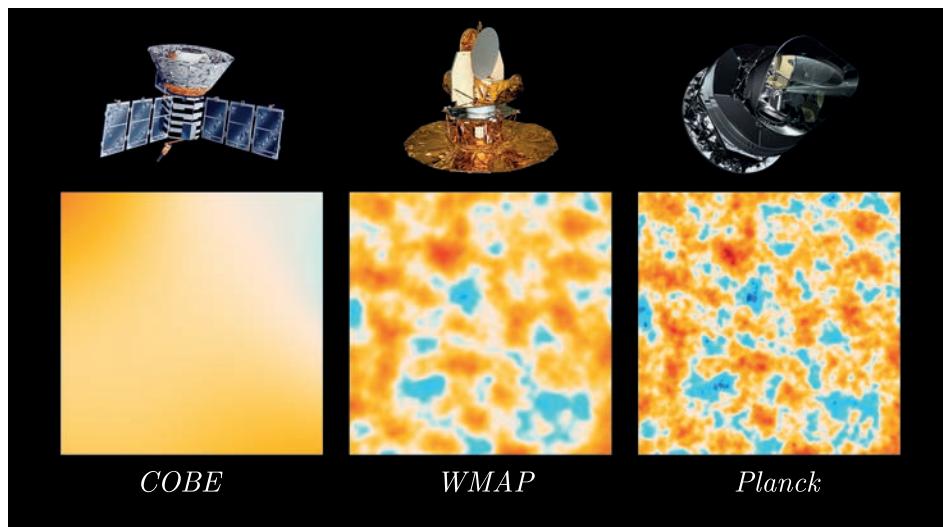


(c)

Figure 6.2 The three ground-based microwave telescopes that are currently in operation at the South Pole: (a) BICEP3, (b) the South Pole Telescope, (c) BICEP Array.

Balloon-borne experiments have also been used to observe small areas of the CMB from very high altitude, where the effects of the Earth's atmosphere on the microwave signal are much smaller.

Ultimately, a shared shortcoming of all ground-based and balloon-borne microwave telescopes is that each can only observe a small fraction of the sky. To image the *entire* sky in microwaves using a *single* telescope, the only option is to launch it into space. Figure 6.3 shows examples of the outputs from three space telescopes that have performed all-sky surveys of the CMB with increasing sensitivity and angular resolution.



This figure shows relatively small 10-degree-square patches that illustrate the progressive improvement afforded by these different angular resolutions. All three telescopes mapped the entire sky at microwave frequencies.

Figure 6.3 Comparison of the improving angular resolution of three space telescopes that have revolutionised our understanding of the Universe. The *COBE*, *WMAP* and *Planck* satellites produced maps with angular resolutions of 7° , 0.25° and 0.08° (5 arcminutes), respectively.

The first of these missions was NASA's *Cosmic Background Explorer* (*COBE*), which was launched in 1989. *COBE* measured the energy distribution of CMB photons, showing that they follow the perfect black-body spectrum shown in Figure 6.4. The measured black-body temperature enables us to make inferences about the conditions in the early Universe, because the radiation temperature evolved as $T \propto 1 + z$ throughout cosmic history (see Chapter 1).

COBE also provided the first conclusive evidence for the existence of small fluctuations in temperature of CMB photons coming from different directions in the sky.

Janskys

Figure 6.4 plots the intensity of the CMB in units of megajanskys per steradian. The jansky (Jy) is a non-SI unit of spectral flux density that is often used by observational cosmologists and astronomers. In SI units, 1 Jy is equivalent to $10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1}$.

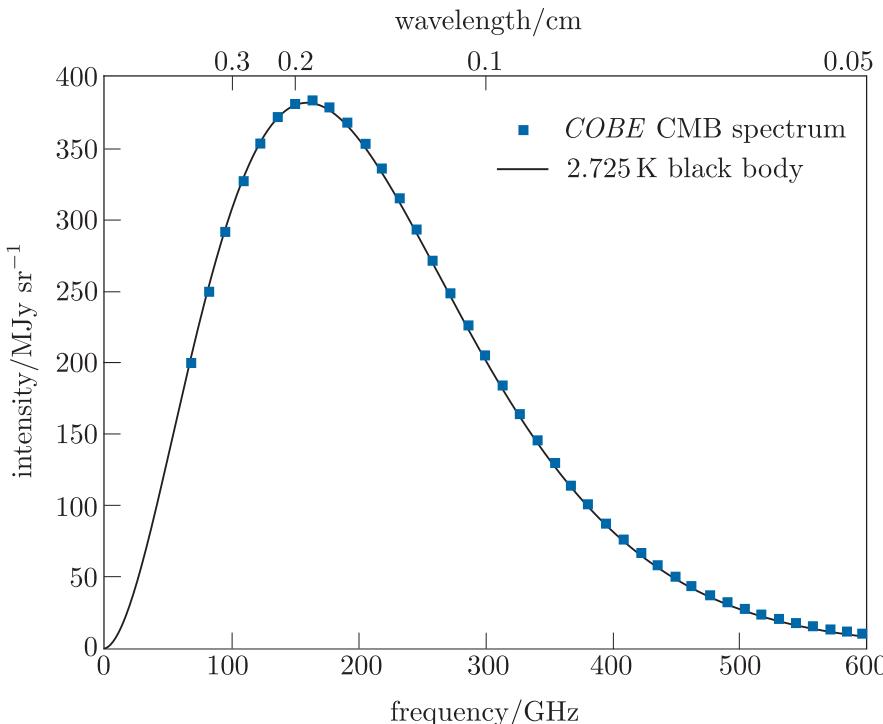


Figure 6.4 The energy spectrum of the CMB as measured by the FIRAS instrument aboard the *COBE* satellite. The curve represents a near-perfect black-body spectrum with a characteristic temperature $T_{\text{CMB}} = 2.725 \text{ K}$.

- In Chapter 1 you learned that this black-body spectrum is characteristic of emission from a medium in perfect thermal equilibrium. What does this equilibrium imply about the contents of the early Universe?
- It implies that at some time in its very early history the entire observable Universe was causally connected and able to exchange energy.

COBE was decommissioned in 1993. It was succeeded in 2001 by NASA's *Wilkinson Microwave Anisotropy Probe (WMAP)* and later, in 2009, by the European Space Agency's *Planck* satellite. Both satellites mapped the CMB temperature fluctuations with unprecedented sensitivity and angular resolution. The *Planck* telescope could measure temperature differences as small as one part in a million between patches of sky separated by as little as 5 arcminutes. In the next chapter you will learn how the exquisitely detailed data provided by *WMAP* and *Planck* have allowed cosmologists to derive constraints on the fundamental properties of the Universe.

6.2.2 The noisy microwave sky

Even with sophisticated space telescopes like *WMAP* and *Planck*, making precise measurements of the CMB temperature fluctuations is still difficult. One reason for this is that the CMB is not the only source of microwaves arriving on Earth from space: several other astrophysical processes also emit microwave radiation and can contaminate or obscure the CMB signal. In addition, there are physical phenomena that can distort the observed spectra of CMB photons at different locations on the sky.

The CMB solar dipole

Figure 6.5a shows an all-sky map of the temperature variations in the microwave signal that the *Planck* telescope actually detects. One of the most obvious features of this map is a large-scale temperature gradient called the **CMB solar dipole**.* The pattern of the solar dipole on its own is shown in Figure 6.5b. The solar dipole results from the peculiar motion of the Solar System through space. Fundamentally, it reflects a result from special relativity that observers in different inertial frames will measure different wavelengths for the same photon.

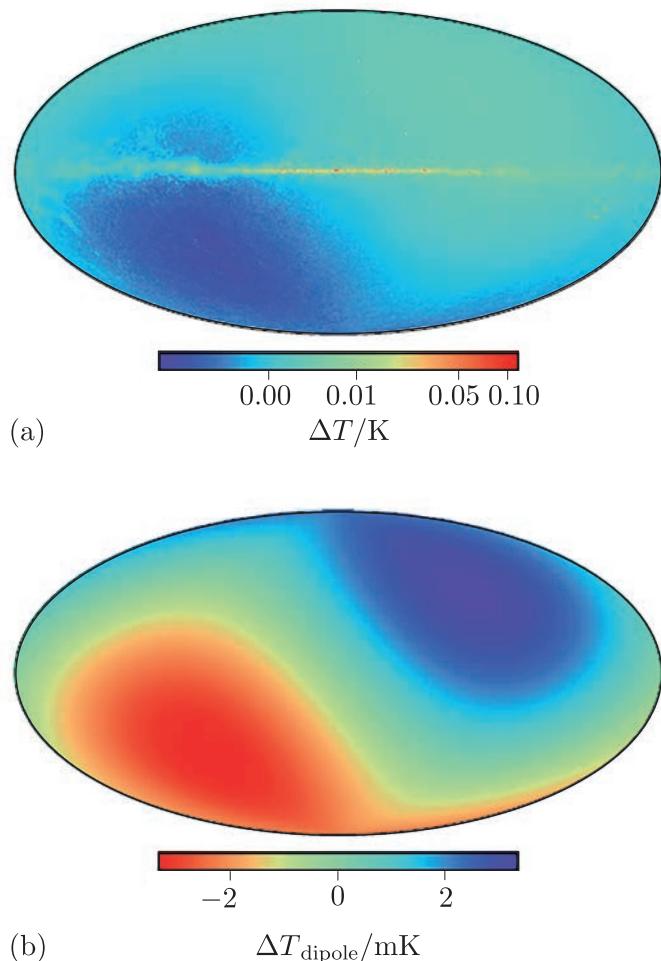


Figure 6.5 (a) An all-sky map of microwave emission coming from space. The CMB solar dipole is clearly visible as large, relatively warm (upper right quadrant) and relatively cool (lower left quadrant) regions. (b) The pattern of temperature variations due to the solar dipole alone.

- Which two inertial frames are relevant when thinking about the origin of the CMB solar dipole?

*The CMB is also modified by another time-varying dipole pattern that is produced by the orbital motion of the Earth around the Sun, but the size of this effect is about 12 times smaller than the solar dipole. The observation strategy of *Planck* is designed so that the signal averages to zero over the course of one year's observations.

- The first inertial frame is that in which the Solar System is instantaneously at rest. The second frame is one in which the CMB appears isotropic, with no dipole pattern. This is often described as the *CMB rest frame*.

In fact, the first inertial frame is only approximately inertial because the Solar System is being *accelerated* in several directions. Recall that inertial frames move with *constant* velocity with respect to each other. The Solar System is orbiting around the centre of the Milky Way and the Milky Way itself is orbiting around the common centre of mass of the **Local Group** (the gravitationally bound group of galaxies of which the Milky Way is a member). Even the Local Group itself is falling under gravity, towards a local overdensity in the cosmic mass distribution.

In the CMB rest frame, the *present* surface of last scattering is at rest. Remember that observers at different locations in the Universe see different surfaces of last scattering so, because the Earth is moving, the surface of last scattering will change slightly from instant to instant.

In Chapter 2 you learned about the Lorentz transformations, which can be used to convert (or ‘map’) the four spacetime coordinates from one inertial frame to another. The Lorentz transformations can also be used to show that the motion of the Solar System relative to the CMB rest frame produces a Doppler shift in the frequencies of CMB photons that are observed from Earth or from Earth orbit.

To express this Doppler shift mathematically we need to specify the relative motion of our two inertial frames more formally. We will choose our coordinate axes so that the Solar System rest frame (which we will label as S) and the CMB rest frame (which we will label as S') are in the ‘standard’ configuration that you saw in Figure 2.1, with relative velocity V parallel to their common x -axis.

- Does making this specific choice of coordinate axes mean that our Doppler shift expression will not be generally applicable?
- No. We are only considering two inertial frames, so it is always possible to choose a coordinate system that has its x -axis parallel to the direction of their relative motion.

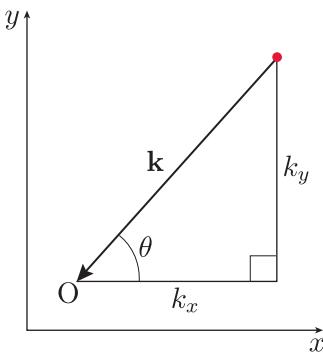


Figure 6.6 An observer at O who is at rest in the inertial frame S detects a photon propagating in the xy -plane. The vector \mathbf{k} defines the direction of the photon’s propagation and is shown by the arrow. The photon’s trajectory makes an angle θ with the x -axis.

Now consider the scenario shown in Figure 6.6, in which an observer at rest in frame S detects a photon with frequency ν . The vector \mathbf{k} , with components k_x and k_y , defines the direction of the photon’s propagation in S, and the angle between \mathbf{k} and the x -axis is θ . An observer at rest in S' will measure a different value ν' for the photon’s frequency. If γ is the Lorentz factor corresponding to the relative velocity of S and S', then ν and ν' are related by the relativistic Doppler-shift formula

$$\nu' = \gamma\nu \left(1 - \frac{V}{c} \cos\theta\right) \quad (6.1)$$

Transforming between S and S' produces a redshift if k_x is parallel to the relative velocity V of the two frames ($\cos\theta > 0$), and a blueshift if

k_x and V are antiparallel ($\cos \theta < 0$). These frequency shifts preserve the *shape* of the CMB's black-body spectrum but modify its apparent *temperature*. The change in apparent temperature is directly proportional to the change in measured photon frequency, so if T and T' are the CMB temperatures measured in S and S', respectively, then Equation 6.1 can be rewritten as:

$$T' = \gamma T \left(1 - \frac{V}{c} \cos \theta \right) \quad (6.2)$$

To show that this expression implies the dipole pattern that is observed, we can first assume that the CMB temperature in its rest frame T' is constant in all directions. Now, by replacing γ with its definition in terms of V (Equation 2.5), we can write a series expansion for the *fractional* temperature offset induced by the motion of the Solar System:

$$\begin{aligned} \frac{T}{T'} &= \left(1 - \frac{V^2}{c^2} \right)^{1/2} \left(1 - \frac{V}{c} \cos \theta \right)^{-1} \\ &= \left(1 - \frac{1}{2} \frac{V^2}{c^2} + \dots \right) \left(1 + \frac{V}{c} \cos \theta + \frac{V^2}{c^2} \cos^2 \theta + \dots \right) \\ &= 1 + \frac{V}{c} \cos \theta + \frac{V^2}{c^2} \left(\cos^2 \theta - \frac{1}{2} \right) + \dots \end{aligned} \quad (6.3)$$

The second term in Equation 6.3 describes the CMB dipole as a *fractional offset* from the mean temperature of the CMB in the rest frame of the surface of last scattering. CMB photons arriving from the direction in which the Solar System is moving are blueshifted and their spectrum appears hotter, while those moving parallel to its motion are redshifted and their spectrum appears cooler. Note that there are also higher-order terms but their impact on the measured CMB temperature turns out to be very strongly suppressed because the Solar System's velocity $V \ll c$.

Example 6.1

The *Planck* satellite measured the amplitude of the solar dipole to be $\Delta T \approx 3.4 \text{ mK}$.

Using this measurement, estimate the speed of the Earth in km s^{-1} relative to the rest frame of the surface of last scattering. You may assume that the *Planck* satellite is at rest in the rest frame of the Solar System and that the Solar System's velocity $V \ll c$. Give your answer to 3 significant figures.

Solution

To solve this problem we will use Equation 6.2. We are told to assume that $V \ll c$, which means that $\gamma \approx 1$. Rearranging gives

$$\frac{T}{T'} \approx \left(1 - \frac{V}{c} \cos \theta \right)^{-1}$$

Now, we are told that ΔT represents the measured amplitude of the solar dipole, and Figure 6.5b shows that the solar dipole pattern is symmetric around $\Delta T = 0$. This means that the measured CMB temperature in the Solar System rest frame varies between $T_{\min} = T' - \Delta T$ and $T_{\max} = T' + \Delta T$.

Let's consider the maximum in the solar dipole[†], which corresponds with blueshifted CMB photons that are propagating antiparallel to the motion of the solar system, so $\cos \theta = -1$. That means we can rewrite Equation 6.2 as

$$\frac{T_{\max}}{T'} = \frac{T' + \Delta T}{T'} \approx \left(1 + \frac{V}{c}\right)^{-1}$$

Now, you learned in Chapter 1 that the mean temperature of the CMB in its rest frame is $T' \approx 2.7\text{ K}$ and we are told in the question that the dipole amplitude $\Delta T \approx 3.4 \times 10^{-3}\text{ K}$. Rearranging our previous equation and using these numerical values, we can write:

$$\begin{aligned} V &= c \left(\frac{T'}{T' + \Delta T} - 1 \right) \\ &= 3 \times 10^5 \text{ km s}^{-1} \times \left(\frac{2.7 \text{ K}}{2.7 \text{ K} + 3.4 \times 10^{-3} \text{ K}} - 1 \right) \\ &= -378 \text{ km s}^{-1} \end{aligned}$$

This means that the speed of the Earth relative to the rest frame of the surface of last scattering is $\sim 378 \text{ km s}^{-1}$.

- Look again at the short CMB introduction in Chapter 1. How does the amplitude of the solar dipole ($\Delta T \approx 3.4 \text{ mK}$) compare to the maximum amplitude of the CMB temperature fluctuations?
- The maximum amplitude of CMB fluctuations is $\lesssim 300 \mu\text{K}$ (see the scale provided in Figure 1.9). The solar dipole amplitude is ten times larger than this value.

The variation in observed sky temperature produced by the solar dipole is large enough to significantly distort the observed shapes and amplitudes of the CMB temperature fluctuations. As you will learn later in this chapter, these distortions would effectively prevent accurate inference of the cosmological parameters using CMB observations. Fortunately, cosmologists have developed sophisticated techniques to remove the solar dipole signal from maps of the CMB.

[†]The solar dipole is symmetric so we could also have chosen to consider the solar dipole minimum and we would arrive at the same result.

Astrophysical foreground emission

Figure 6.7 shows the *Planck* skymap after subtracting the solar dipole signal, but before removing the astrophysical microwave foreground. The map shows a bright band of microwave emission coinciding with the Galactic plane, as well as several microwave-emitting, filamentary structures that extend to higher Galactic latitudes.

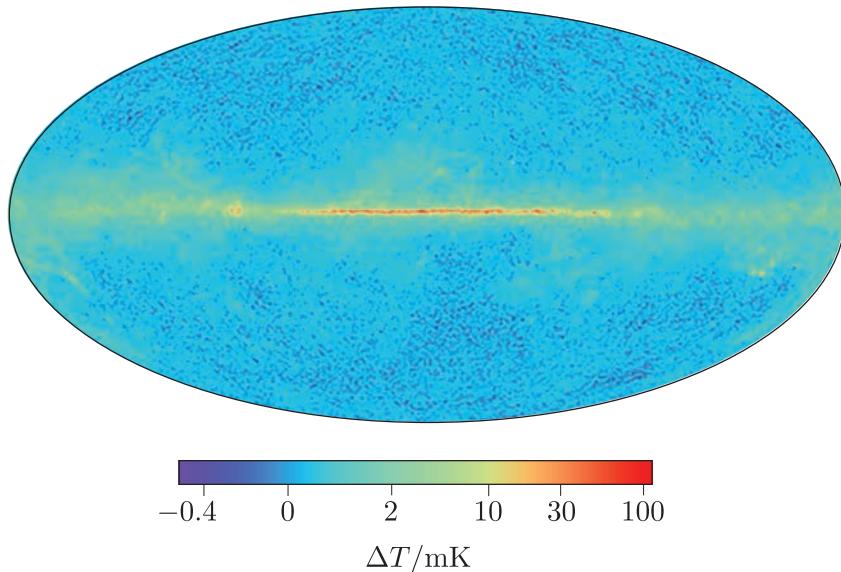


Figure 6.7 *Planck* all-sky map after correcting for the distortion of the solar dipole, but before removing the astrophysical microwave foreground.

Much of this microwave emission comes from radiating particles in the Milky Way. These particles are much closer than the surface of last scattering, so the microwaves they emit are often referred to as ‘Galactic foreground’ emission. Several physical mechanisms contribute to the Galactic foreground:

- Relativistic electrons emit microwaves via **synchrotron radiation** as they follow spiralling trajectories in the Galaxy’s magnetic fields.
- Microwaves in the form of **bremsstrahlung radiation** are emitted when free electrons are accelerated in the Coulomb fields of ions in the **interstellar medium** (ISM). This is commonly known as *free-free emission*.
- Excited molecules in giant molecular clouds and the ISM emit microwaves at specific frequencies as they decay back towards their ground states.
- Interstellar dust grains that rotate at particular frequencies can radiate microwaves if they have a non-zero **electric dipole moment**.
- *Thermal* emission from dust grains at a similar temperature to the CMB is, perhaps unsurprisingly, also in the microwave frequency range.

If the foreground emission were uniform then it would be easy to remove it and reveal the CMB signal. However, the foreground changes by large amounts on different angular scales, which can mask or distort the CMB temperature fluctuations that cosmologists are trying to measure.

Figure 6.8 plots the amount of variation that is produced at different microwave frequencies by the various foreground emission processes and compares this variability with the typical sizes of CMB temperature fluctuations observed by the *Planck* satellite. At almost all frequencies, fluctuations in the combined foreground emission completely dominate those in the CMB signal. The variations are quantified in terms of the **root mean square** (RMS) of the corresponding foreground component's signal around its mean level. The dashed lines show the total RMS temperature variation that is contributed by all the foreground emission processes. At all frequencies except ~ 90 GHz this total variation exceeds that of the CMB itself, which is shown by the thin red line.

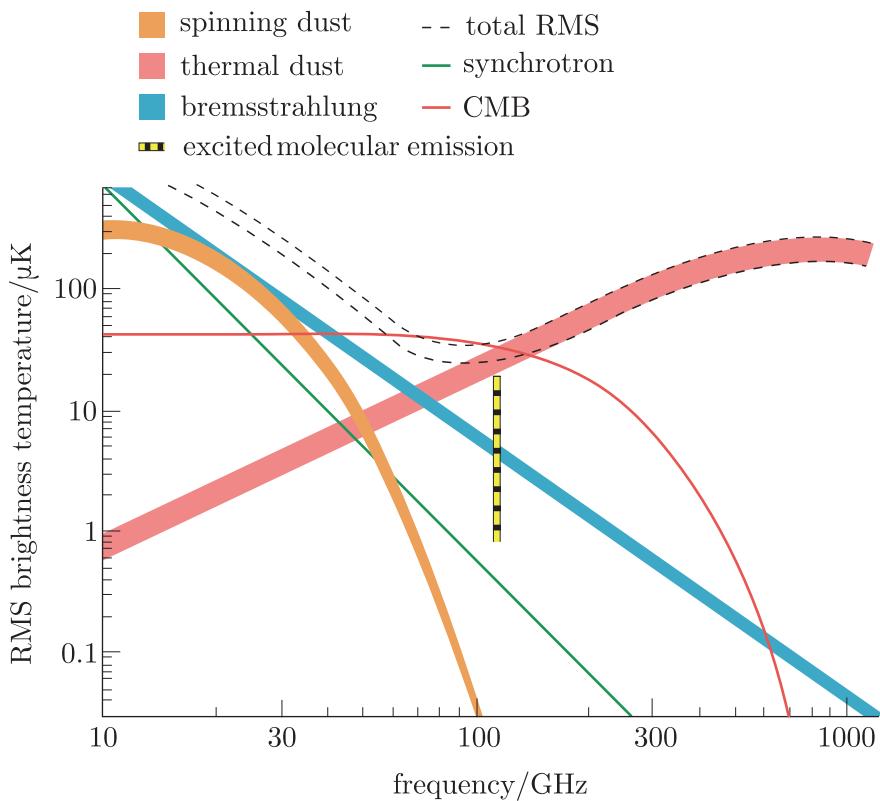


Figure 6.8 The size of *variations* in the measured effective temperature of microwave emission from different components of the diffuse Galactic foreground emission.

Fortunately, cosmologists are able to use observational data and theoretical knowledge about the physical processes involved to build detailed models for the emission from each foreground component as functions of both sky location and photon frequency. By subtracting the combined emission that

is predicted by these models, cosmologists are able to recover the map of the CMB temperature fluctuations shown in Figure 6.9, which is effectively free from foreground contamination.

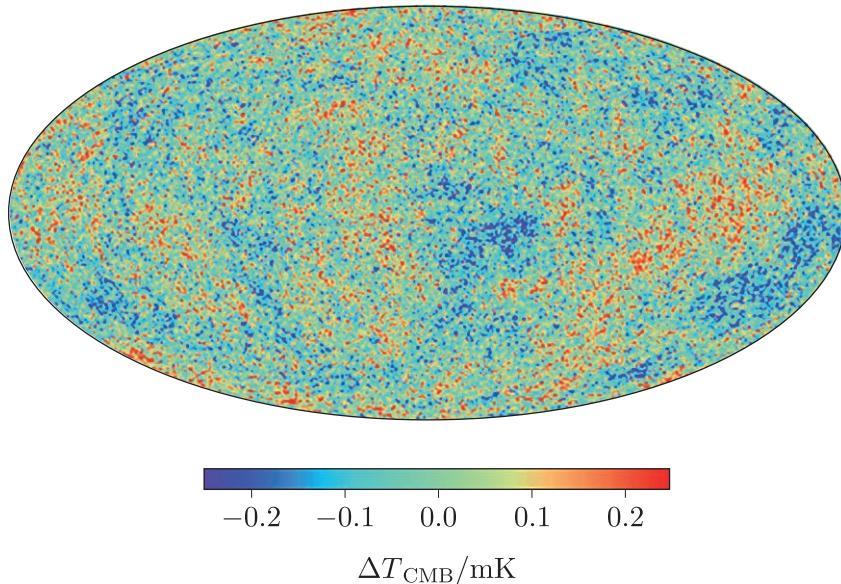


Figure 6.9 The CMB temperature fluctuations after removal of foreground contaminants and the solar dipole. The remaining temperature fluctuations are of the order of hundreds of μK , but they encode a wealth of information about the early Universe and the cosmological parameters.

Next you will read about how cosmologists can mathematically describe and statistically model the pattern of temperature fluctuations shown in Figure 6.9. This mathematical description will allow us to effectively listen to the acoustic oscillations that were introduced in Section 6.2, and, ultimately, to constrain several of the cosmological parameters.

6.3 CMB temperature fluctuations

In this section you will learn more about the CMB temperature fluctuations that were discovered and then probed in detail by the *COBE*, *WMAP* and *Planck* satellites. We will start by constructing a mathematical framework to describe individual temperature fluctuations before introducing the concept of the CMB angular power spectrum, which is a mathematical tool that can be used to examine the amplitudes of sets of fluctuations on different angular scales. In the next chapter you will then learn how various features of the power spectrum arise from processes that were operating during the early Universe, up until the epoch of last scattering.

6.3.1 Describing individual temperature fluctuations

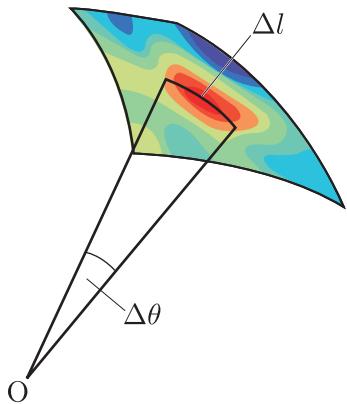


Figure 6.10 An individual CMB temperature fluctuation. An observer at O measures the angular size of the fluctuation to be $\Delta\theta$. Its physical size at the epoch of last scattering is Δl .

For the purposes of this section, let's loosely define an *individual* temperature fluctuation as a contiguous region of the CMB that has a particular observed temperature and is bounded by other regions that have different temperatures.

Before we consider the properties of the CMB as a whole, let's start by exploring two observable properties that can be used to characterise an individual CMB temperature fluctuation like the one shown in Figure 6.10: its *angular size* and its *amplitude*. The angular size $\Delta\theta$ of a temperature fluctuation is interesting because it is related to the physical size Δl of a contiguous region of the Universe at the epoch of last scattering that contained fluids with similar physical properties like density, temperature and velocity.

- Which quantity that you learned about in Chapter 5 relates Δl and $\Delta\theta$ to each other?
- They are related by the angular diameter distance to the surface of last scattering $d_A(z_{ls})$ such that $\Delta l = d_A(z_{ls}) \Delta\theta$.

Let's estimate the physical size of the smallest fluctuations that were resolved by the *Planck* satellite. Their angular size was at the very limit of *Planck*'s angular resolution, $\Delta\theta \approx 5$ arcminutes.

Example 6.2

In Chapter 5 you learned that the present horizon distance $d_{\text{hor}}(t_0)$ is equivalent to the present proper distance that a photon has travelled since the big bang. For the real Universe, it can be shown that $d_{\text{hor}}(t_0) = 14\,165$ Mpc.

- Using this fact, show that when the redshift of an observed object is very large, such that $z \rightarrow \infty$, then a good approximation for its angular diameter distance is

$$d_A \approx \frac{d_{\text{hor}}(t_0)}{z} \quad (6.4)$$

- Use this approximation to compute the physical size of the smallest temperature fluctuation that the *Planck* telescope was able to resolve on the surface of last scattering. How does this length scale compare with the physical sizes of objects in the Universe today?

Solution

- Because any observable object must be within the particle horizon, the fact that $d_{\text{hor}}(t_0)$ is finite means that the present proper distance to any such object must also remain finite. In fact, it must converge to $d_{\text{hor}}(t_0)$ as the object's redshift $z \rightarrow \infty$.

The angular diameter distance is related to the present proper distance by Equation 5.15:

$$d_A = \frac{d_p(t_0)}{1+z} \quad (\text{Eqn 5.15})$$

When $z \rightarrow \infty$, we can make the approximations that $1+z \approx z$ and that $d_p(t_0) \approx d_{\text{hor}}(t_0)$, so Equation 5.15 simplifies to the required solution:

$$d_A \approx \frac{d_{\text{hor}}(t_0)}{z}$$

- (b) Using Equation 6.4 and the value of $d_{\text{hor}}(t_0) = 14\,165 \text{ Mpc}$ provided in the question, the approximate angular diameter distance to the surface of last scattering is

$$d_A(z_{\text{ls}}) = \frac{d_{\text{hor}}(t_0)}{z_{\text{ls}}} \approx \frac{14\,165 \text{ Mpc}}{1090} \approx 13 \text{ Mpc}$$

Using this result we can write an expression for the physical size, l , of a CMB fluctuation on the surface of last scattering that subtends an angle $\Delta\theta$ on the sky today:

$$\begin{aligned} l &= d_A(z_{\text{ls}}) \left(\frac{\Delta\theta}{\text{radians}} \right) \approx 13 \text{ Mpc} \left(\frac{\Delta\theta}{\text{radians}} \right) \\ &\approx \frac{\pi}{60 \times 180} \times 13 \text{ Mpc} \left(\frac{\Delta\theta}{\text{arcminutes}} \right) \approx 3.8 \text{ kpc} \left(\frac{\Delta\theta}{\text{arcminutes}} \right) \end{aligned}$$

Planck was able to resolve angular structures as small as 5 arcminutes. These correspond to regions of the Universe at last scattering with physical proper sizes $\approx 19 \text{ kpc}$. In the present-day Universe, length scales of $\sim 20 \text{ kpc}$ are comparable to the physical size of galaxies like the Milky Way.

This is a remarkably good approximation. Evaluating Equation 5.9 numerically and using the result in Equation 5.15 gives $d_A(z_{\text{ls}}) = 12.77 \text{ Mpc}$.

Exercise 6.1

If the proper size of a region at z_{ls} is 19 kpc, what is its proper size today?

Example 6.3

In Example 6.2 you saw that the angular diameter distance for $z \approx z_{\text{ls}}$ can be well approximated in terms of the current horizon distance $d_{\text{hor}}(t_0)$.

In this short example you will explore a piece of Python code to compute $d_A(z_{\text{ls}})$ directly using Equation 5.15. This will involve first numerically evaluating the integral in Equation 5.9 to find $d_p(t_0)$.

Solution

Please refer to the Example 6.3 Jupyter Notebook in the online module resources to see the solution to this example.

To describe the *amplitude* of CMB temperature fluctuations, cosmologists define a dimensionless quantity that describes the relative difference between the CMB temperature $T(\theta, \phi)$ measured in the direction defined by the angular spherical coordinates θ and ϕ , and the mean CMB temperature $\langle T \rangle$, averaged over the whole sky.

The amplitude of CMB temperature fluctuations

$$\frac{\Delta T}{T}(\theta, \phi) \equiv \frac{T(\theta, \phi) - \langle T \rangle}{\langle T \rangle} \quad (6.5)$$

$\langle T \rangle$ can be calculated by evaluating a surface integral:

$$\langle T \rangle = \frac{1}{4\pi} \int_0^\pi \int_0^{2\pi} T(\theta, \phi) \sin \theta \, d\theta \, d\phi$$

- What is the mean value of $\Delta T/T$ for the whole sky? If $\Delta T/T$ is positive in a particular direction, what does that tell you about the CMB temperature at that location on the sky? What if $\Delta T/T$ is negative?
- By construction, the mean value of $\Delta T/T$ for the whole sky is zero. $\Delta T/T$ will be positive for points on the sky that are warmer than average and negative for points that are cooler than average.

The pattern of temperature fluctuations shown in Figure 6.9 is very complex. It might look like random noise, but in fact it comprises multiple overlapping components with a large range of angular sizes. Cosmologists are interested in the relative prevalence of differently sized fluctuations because this encodes a range of important information about the cosmological parameters and the physics of the early Universe.

To separate out and study sets of fluctuations on different angular scales, it is common practice to expand $\Delta T/T$ into a sum of periodic functions called **spherical harmonics**. The spherical harmonics $Y_{lm}(\theta, \phi)$ are two-dimensional functions expressed in terms of the spherical coordinates θ and ϕ that can be used to decompose functions that are defined on the surface of a sphere. Spherical-harmonic expansion is closely analogous to **Fourier expansion**, which can be used to decompose arbitrary functions into a weighted sum of sines and cosines. More detailed background information about spherical harmonics is provided in the next section.

Spherical harmonics

If the integral of the square of a function f between $-\infty$ and $+\infty$ is finite, then f is said to be a square-integrable function. Any square-integrable function $f(\theta, \phi)$ that is defined using spherical coordinates θ, ϕ can be decomposed into an infinite sum of terms:

$$f(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta, \phi) \quad (6.6)$$

The term $a_{\ell m}$ are a set of numerical coefficients that will be explained shortly. The functions $Y_{\ell m}(\theta, \phi)$ are called spherical harmonics and are defined, for integer values of $\ell > 0$ and $-\ell \leq m \leq \ell$, by the expression

$$Y_{\ell m}(\theta, \phi) = \sqrt{\frac{2\ell+1}{4\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_{\ell}^m(\cos \theta) e^{im\phi}$$

Figure 6.11 (overleaf) illustrates all the spherical harmonics for which $\ell \leq 3$. Each $Y_{\ell m}$ defines a different two-dimensional, periodic pattern on the surface of a sphere and the values of ℓ and m determine the number of oscillations in the θ - and ϕ -directions, respectively.

In the figure, each oval shows an **Aitoff projection** of the entire surface of a sphere, with the 2π radians around its equator shown in the horizontal (ϕ) direction, and the π radians between its two poles shown in the vertical (θ) direction. Many all-sky maps of the CMB, including the one shown in Figure 6.9, are also drawn using the Aitoff projection. Red and blue colours indicate positive and negative values, respectively. Zero-valued nodes of the $Y_{\ell m}$ function are shown in white.

The dependence of $Y_{\ell m}$ on θ is completely specified by the function $P_{\ell}^m(\cos \theta)$, which represents an **associated Legendre polynomial**. The integer ℓ is often referred to as the ‘multipole number’ and, together with m , it determines the number of *nodes* at which $P_{\ell}^m(\cos \theta) = 0$ between the poles of the sphere, for $-1 < \cos \theta < 1$.

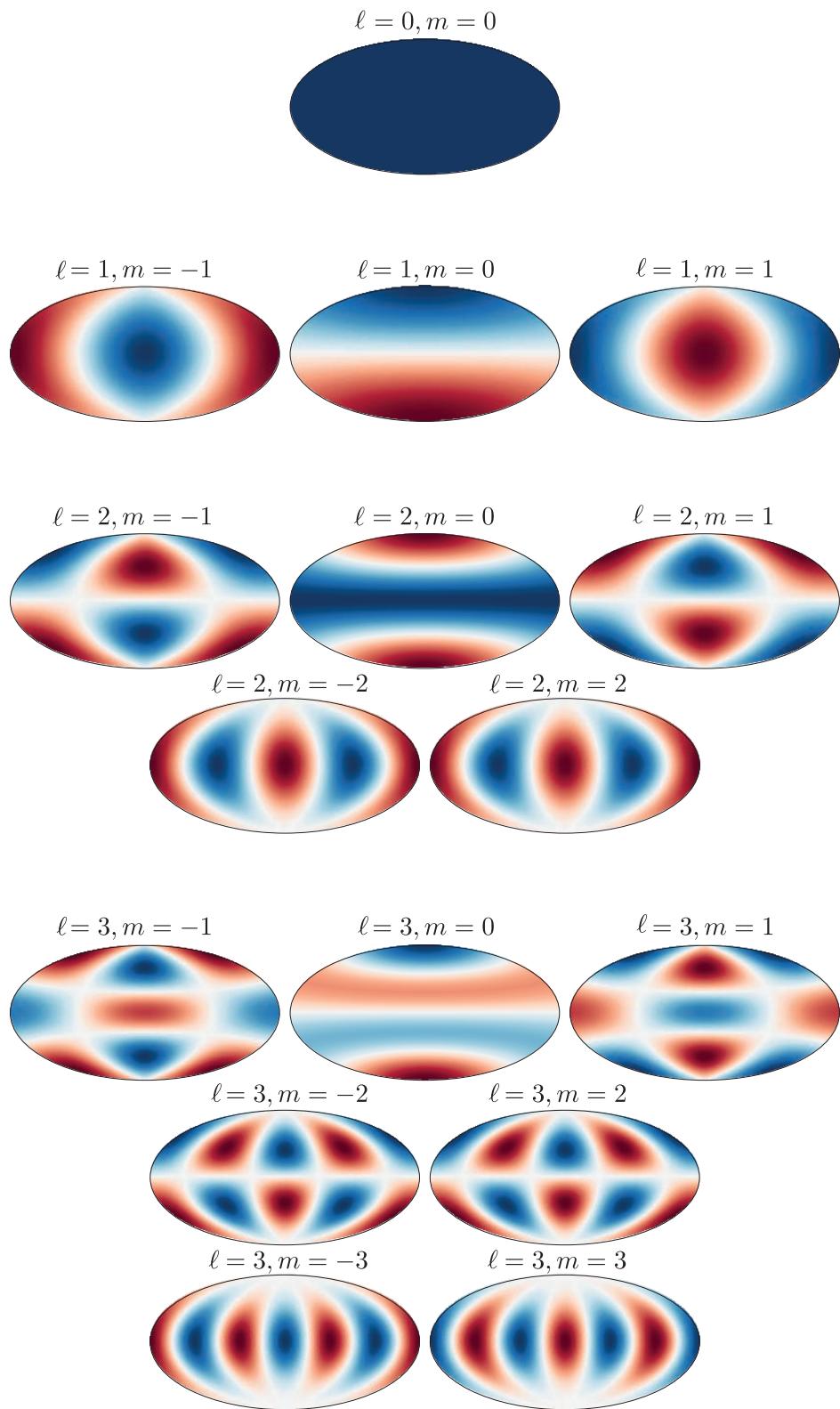


Figure 6.11 An illustration of the real part of the spherical harmonics for $\ell \leq 3$.

Legendre Polynomials

The formula needed to evaluate $P_\ell^m(\cos \theta) = 0$ is a little complicated and we will not quote it here. Instead, Figure 6.12 shows the associated Legendre polynomials for $0 \leq \ell \leq 5$ and $0 \leq m \leq 1$ to give an idea of their shapes. For $\ell + m > 1$, the associated Legendre polynomials are oscillatory functions with $\ell - |m|$ nodes (where $P_\ell^m = 0$) excluding the end points, and $\ell - |m| + 1$ extrema (where $dP_\ell^m/dx = 0$) between -1 and 1 .

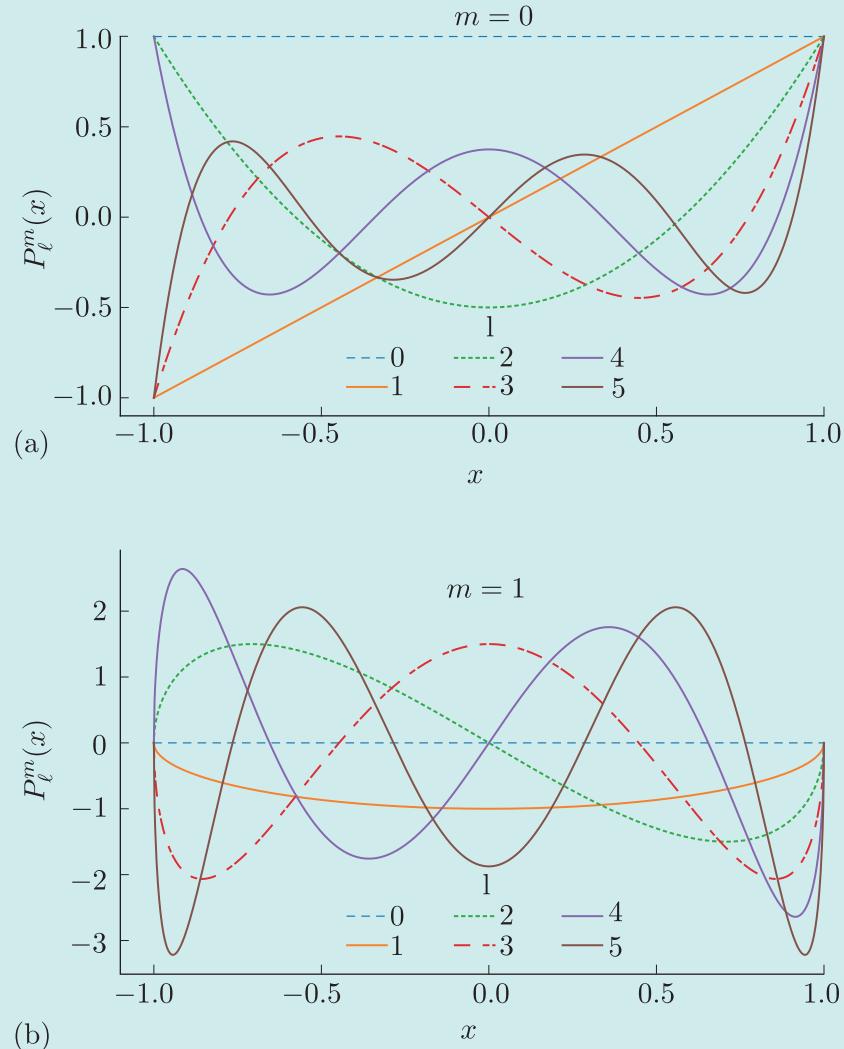


Figure 6.12 The associated Legendre polynomials for $0 \leq \ell \leq 5$ and $0 \leq m \leq 1$.

The azimuthal (ϕ) dependence of $Y_{\ell m}$ is a periodic function, $e^{im\phi}$, that wraps around the circumference of the sphere with $|m|$ complete periods (or, equivalently, $2|m|$ nodes) in the interval $0 \leq \phi \leq 2\pi$.

The coefficients $a_{\ell m}$ of the spherical harmonics in Equation 6.6 are complex numbers in general, and can be evaluated using

$$a_{\ell m} = \int_{-\pi}^{\pi} \int_0^{2\pi} f(\theta, \phi) Y_{\ell m}^*(\theta, \phi) d\Omega$$

where $d\Omega = \sin \theta d\theta d\phi$ and $Y_{\ell m}^*(\theta, \phi)$ is the complex conjugate of $Y_{\ell m}(\theta, \phi)$.

Having summarised the mathematics of spherical harmonics, we can now use them to write an expansion for the complex pattern of CMB temperature fluctuations as a sum of simpler quasi-periodic functions:

$$\frac{\Delta T}{T}(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta, \phi) \quad (6.7)$$

Individual spherical harmonics are often referred to as *modes* in reference to analogous vibrational modes in acoustics.

It may help to remind yourself that there are 2π radians around the equator of a sphere, but only π radians between its two poles.

Here we have chosen $\Delta T/T$, the observable pattern of temperature fluctuations, to be the function $f(\theta, \phi)$ in Equation 6.6 that we wish to decompose. Each spherical harmonic $Y_{\ell m}$ in the sum represents fluctuations with a particular angular scale ($\Delta\theta \approx \pi/\ell$ radians) and is weighted by an amplitude coefficient $a_{\ell m}$, which carries information about the relative contribution to $\Delta T/T$ from fluctuations on that particular angular scale. The correspondence between ℓ and $\Delta\theta$ is most obvious for the $m = 0$ spherical harmonics illustrated in Figure 6.11, but a closer inspection of the $m \neq 0$ harmonics should convince you that each contiguous region of positive or negative values subtends a solid angle that is characterised by a linear angular scale $\sim \pi/\ell$ radians.

Online resources: exploring the spherical harmonics

The next week-long Python activity, to be studied after completing Chapter 7, will give you the opportunity to explore spherical harmonics in a hands-on way, and learn more about the ways that they can be used to understand the CMB temperature fluctuations.

6.3.2 Statistical descriptions of the CMB

In Chapter 1 you learned that the CMB temperature fluctuations reflect the existence of local inhomogeneities in the matter and radiation density of the Universe at the epoch of last scattering. In later chapters you will learn more about the physical processes that generated these density fluctuations even earlier in the history of the Universe, when $z \gg z_{ls}$. In this chapter we will not need to understand the fundamental cause of these density perturbations. Instead, we will focus on modelling how they should *evolve* and grow in the time before the epoch of last scattering.

Instead of assuming a *particular* physical origin for the CMB temperature fluctuations, we can model them mathematically as having been generated by a random statistical process. We can then treat the particular pattern of fluctuations that we *do* see as just one possible randomly generated realisation. If we could somehow restart the Universe over and over again,

we would generate a new CMB with a different pattern of temperature fluctuations each time. Even if each new Universe had exactly the same cosmological parameters, the random selection of the initial pattern of density perturbations would still result in a new pattern of temperature fluctuations. Our CMB generator is a random process, so we cannot predict the pattern of temperature fluctuations for any *particular* realisation. However, we *can* predict and write down what the statistical *expectation values* are for different observable properties of the CMB. When cosmologists use theoretical models to predict the statistical properties of the CMB, they are actually predicting these expectation values using different choices for the cosmological parameters.

In reality, there is only one CMB that we can observe to measure its properties; our CMB results from one *particular* set of initial density perturbations and it corresponds with the *particular* surface of last scattering that has the Earth at its centre. Observers at different locations in the Universe will observe different CMBs that have a different pattern of temperature fluctuations. Statistically, each of these observed CMBs is a single *sample* from the true pattern of temperature fluctuations that existed throughout the Universe at the epoch of last scattering.

To compare the *observed* properties of our CMB with model predictions, cosmologists need to find some way to estimate those properties' true values using the limited information from the single statistical sample that our particular view of the CMB provides. To do this, they use mathematical functions called **estimators** that can compute theoretical expectation values based on limited statistical samples.

The next section will introduce one of the most important statistical properties of the CMB; you will learn how the theoretical expectation value of this property is defined and how it can be estimated using observational data.

6.3.3 The CMB angular power spectrum

One of the most cosmologically informative properties of the CMB is the variance of $\Delta T/T$, measured on different angular scales. To quantify this variance, cosmologists use a statistical function called the **angular power spectrum**, denoted as C_ℓ .

For a particular value of the multipole number ℓ , the statistical expectation for C_ℓ can be defined as

$$C_\ell \equiv \langle |a_{\ell m}|^2 \rangle$$

where $|a_{\ell m}|^2 = a_{\ell m} a_{\ell m}^*$ is the squared amplitude associated with a particular $Y_{\ell m}$ in the spherical-harmonic expansion of $\Delta T/T$ (Equation 6.7).

C_ℓ represents a *theoretical* average of $|a_{\ell m}|^2$ over an infinite number of randomly generated CMB realisations for model universes that have identical cosmological parameters.

You can think of the expectation value $\langle X \rangle$ of property X as the average that you would calculate if you measured that property's actual value for an infinite number of different random realisations.

Recall that for a particular value of ℓ there are only $2\ell + 1$ spherical harmonics $Y_{\ell m}$ and associated amplitudes $a_{\ell m}$ defined for $-\ell \leq m \leq \ell$.

We use the hat notation \widehat{C}_ℓ here to indicate a statistical estimate for the unknown true value of the quantity C_ℓ .

A very important property of C_ℓ is that its value depends only on the angular scale of fluctuations represented by $Y_{\ell m}$ and not on where those fluctuations appear on the sky. Mathematically, this means that for a particular multipole number ℓ , the corresponding amplitudes $a_{\ell m}$ are *statistically independent* and the theoretical expectation value of C_ℓ is identical for all permitted values of m .

We can use this fact to derive an estimator for the angular power spectrum. The sample of $|a_{\ell m}|^2$ values we observe is unlikely to equal the corresponding set of expected values because of the inherently random nature of the way the CMB temperature fluctuations were produced.

However, once we have expanded $\Delta T/T$ using spherical harmonics, we will have $2\ell + 1$ measured $a_{\ell m}$ values for each value of ℓ . Then, for each value of ℓ , we can use the mean of the squares of all the corresponding measured amplitude values as an estimator, \widehat{C}_ℓ , for the true value of the power spectrum:

$$\widehat{C}_\ell = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 \quad (6.8)$$

Cosmic variance

Intuitively, we might expect that averaging together more measurements of $|a_{\ell m}|^2$ would result in a more accurate estimate, and that is indeed the case. However, no matter how carefully we measure the angular structure of the CMB we observe, the values of \widehat{C}_ℓ that we calculate can never be perfectly accurate estimates of the true values of C_ℓ . Based on fundamental statistical theory, the expected variance of our estimated \widehat{C}_ℓ values around the true values of C_ℓ is

$$\sigma^2(\widehat{C}_\ell, \ell) = \frac{2}{2\ell + 1} \widehat{C}_\ell^2 \quad (6.9)$$

Equation 6.9 represents an irreducible uncertainty in our measurements of the power spectrum, which cosmologists call **cosmic variance**. It reflects the fact that observing from a single location in the Universe provides incomplete information about the underlying distribution of temperature fluctuations that are only *sampled* by the CMB. Cosmic variance is particularly relevant on large angular scales, which correspond to small values of ℓ and therefore have fewer independent $|a_{\ell m}|^2$ values to average.

From theory to observation

Figure 6.13 presents the CMB angular power spectrum measured by the *Planck* satellite. Instead of plotting C_ℓ directly, it is conventional to use a scaled quantity that has units of squared temperature:

$$\Delta_T^2 = \frac{\ell(\ell + 1)}{2\pi} C_\ell \langle T \rangle^2 \quad (6.10)$$

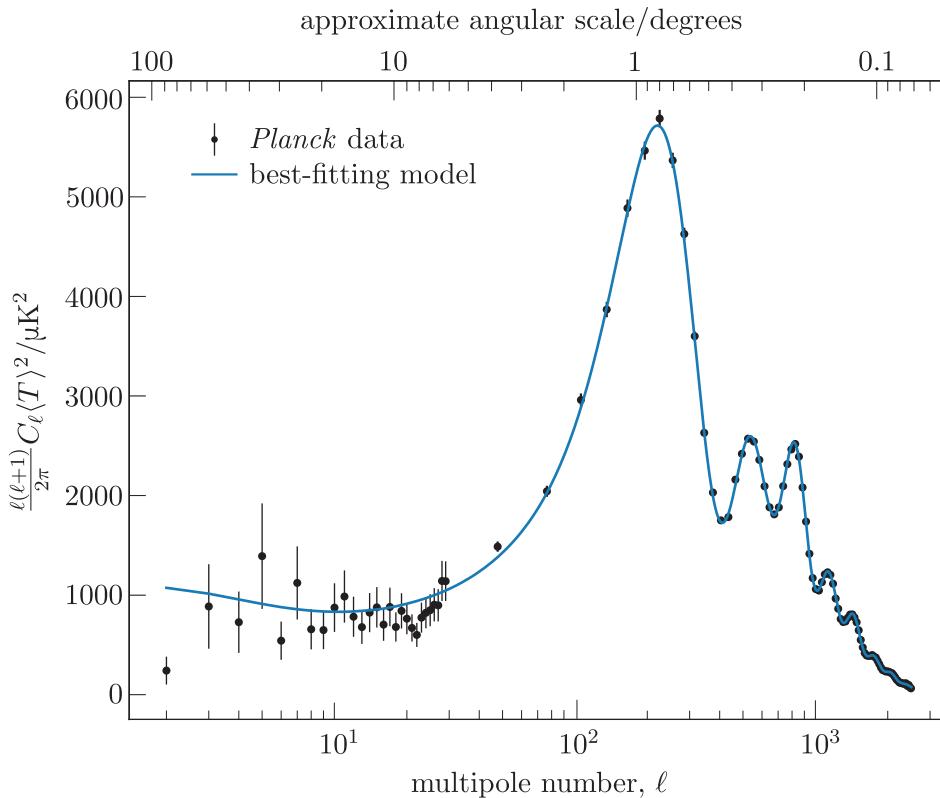


Figure 6.13 The CMB angular power spectrum measured by the *Planck* satellite. The black points show the measurements and the vertical bars show the associated errors.

Note how the effect of cosmic variance increases the scatter on large angular scales (bottom left of figure). The blue line shows the theoretical model that best fits the observed data. The normalisation factor of $\ell(\ell+1)/2\pi$ in Δ_T^2 is a historical convention. It has the useful property that a graph of Δ_T^2 versus ℓ would be approximately flat if the variance of the *density* fluctuations that imprinted their distribution on the CMB was equal on all *physical* scales.

Figure 6.13 shows us that the *observed* spectrum of Δ_T^2 is definitely *not* flat except at very low values of ℓ . It contains several obvious peaks and even the narrow ‘plateau’ that appears for $\ell \lesssim 20$ is not perfectly level. However, this complicated structure in the observed CMB angular power spectrum does not necessarily mean that the *primordial* density fluctuations did not have equal variance on all physical scales, only that this was no longer true by the time of last scattering. If the earliest density perturbations had equal variance on all physical scales, then the observed angular power spectrum tells us that the power spectrum of those initial perturbations was modified by physical processes that operated during the early history of the Universe when $z > z_{ls}$. In the next chapter you will learn what these processes were, how they depend on the values of various cosmological parameters and how we can use the observed power spectrum to constrain what the values of those cosmological parameters are.

6.4 Summary of Chapter 6

- For the first $\sim 370\,000$ years of its history, the baryonic content of the Universe existed as an almost completely ionised plasma. During this time, the Universe was opaque to photons, which could only travel short distances before scattering from free electrons in the plasma.
- The Universe became transparent to photons during a short time interval that ended at the **epoch of last scattering**. The cosmic microwave background (CMB) comprises photons that were released during this interval and have propagated without scattering ever since.
- The CMB photons that we observe today have all travelled very similar distances to reach the Earth. This means that they can be considered to come from a **surface of last scattering** at a redshift $z \approx 1090$.
- The energy spectrum of the CMB photons follows a perfect black-body distribution, which is strong evidence that the whole Universe was in thermal equilibrium very early in its history.
- The radiation temperature of the Universe (and hence the black-body temperature of the CMB) evolved since the epoch of last scattering according to $T \propto 1 + z$.
- The black-body temperature of the CMB spectrum $T(\theta, \phi)$, measured at different points on the sky, exhibits tiny variations with amplitudes $\lesssim 300 \mu\text{K}$ around a mean value of $\langle T \rangle = 2.725 \text{ K}$. These variations are called CMB temperature fluctuations and they are expressed mathematically as

$$\frac{\Delta T}{T}(\theta, \phi) \equiv \frac{T(\theta, \phi) - \langle T \rangle}{\langle T \rangle} \quad (\text{Eqn 6.5})$$

- The complex pattern of fluctuations reflects density perturbations that were present in the Universe at the epoch of last scattering. These density perturbations were produced by an ensemble of overlapping sound waves that were propagating through the fluids filling the early Universe. These sound waves are called **acoustic oscillations**.
- Cosmologists decompose the complex two-dimensional pattern of temperature fluctuations into a series of simpler, periodic, two-dimensional functions $Y_{\ell m}(\theta, \phi)$ called **spherical harmonics**, each of which describes fluctuations on a specific angular scale:

$$\frac{\Delta T}{T}(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta, \phi) \quad (\text{Eqn 6.7})$$

- The CMB **angular power spectrum** $C_{\ell}(\ell)$ measures the contribution of fluctuations on a particular angular scale to the overall variance of the CMB temperature. The spectrum exhibits several peaks, indicating larger contributions to the variance by fluctuations on specific angular scales.

- The coefficients $a_{\ell m}$ in the spherical-harmonic decomposition of the *measured* CMB temperature fluctuations can be used to *estimate* the CMB angular power spectrum:

$$\widehat{C}_\ell(\ell) = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 \quad (\text{Eqn 6.8})$$

- The inherent uncertainty associated with measuring $\widehat{C}_\ell(\ell)$ based on the CMB spectrum observed from just one location (Earth) is known as **cosmic variance**.

Chapter 7 Cosmology with the cosmic microwave background

Chapter 1 introduced you to the basic properties of CMB radiation and its importance as a tool for understanding the Universe, and in Chapter 6 you learned more about the observational properties of the CMB and the mathematical methods used to describe and model it. This chapter will focus on the physical processes that produce the observed anisotropies in the CMB and how cosmologists use CMB observations – and measurements of the angular power spectrum in particular – to measure cosmological density parameters such as Ω_m and Ω_b .

Objectives

Working through this chapter will enable you to:

- explain how physical processes operating before the epoch of last scattering produced the different features we observe in the CMB angular power spectrum
- explain how perturbations in the distributions of matter and radiation in the early Universe imprinted small fluctuations in the temperature of the CMB that we can still observe today
- understand how the values of the cosmological parameters impact the observable properties of the CMB
- explain how various CMB properties would change if the cosmological parameters had different values from those we actually measure for the real Universe
- understand the ways in which cosmologists derive cosmological parameter constraints from very precise observations of the CMB
- understand that CMB observations have some limitations as probes to measure the cosmological parameters but these can often be overcome by simultaneously considering observations of other astrophysical phenomena, such as Type Ia supernovae.

7.1 Explaining the CMB angular power spectrum

The CMB power spectrum shown in Figure 7.1 is exactly the same as the one you saw in Figure 6.13. As in the previous instance, the black points show the measurements and their associated errors. Note the effect of cosmic variance in the increased scatter at larger angular scales. As before, the blue line shows the theoretical model that best fits the observed data.

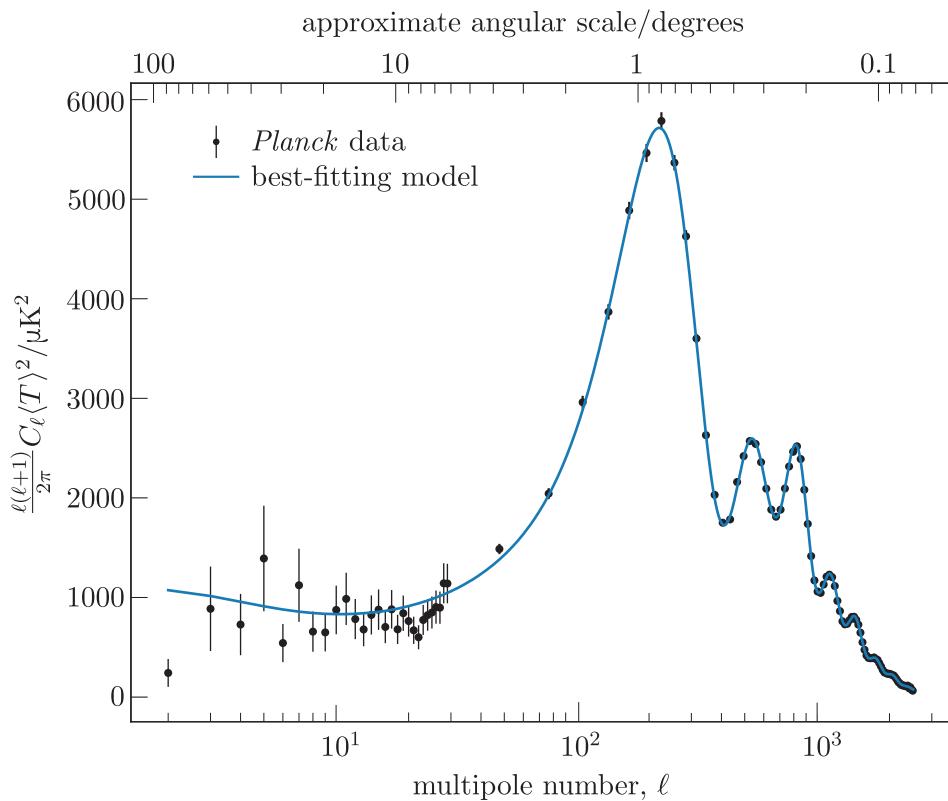


Figure 7.1 The CMB angular power spectrum measured by the *Planck* satellite.

In this section we will examine the structure of the CMB angular power spectrum in much more detail. The different features we investigate will help us to understand the physical processes that were operating before the epoch of last scattering and provide constraints on the values of several cosmological parameters.

To understand why the CMB angular power spectrum exhibits the features that it does, it is important to remember that the *temperature* fluctuations that we observe in the CMB reflect fluctuations in the *density* of the Universe at the epoch of last scattering.

7.1.1 Before the epoch of last scattering

Many cosmological theories predict *primordial* perturbations in the densities of matter and radiation and, further, that these perturbations have equal variance on all physical scales. If this primordial distribution of density perturbations persisted until the epoch of last scattering, then we would expect the values of Δ_T^2 observed in the power spectrum to be equal and independent of the multipole number, ℓ . However, this expectation is not consistent with the complex pattern of peaks in Figure 7.1. Instead, we can see that by the epoch of last scattering, the variance of the density fluctuations had become much larger on some physical size scales while the amplitudes of fluctuations on other scales were correspondingly reduced.

To help us understand how the primordial pattern of density perturbations evolved to produce the CMB angular power spectrum that we *do* observe, we will construct a simplified model for the contents of the early Universe. At the epoch of last scattering, the density parameter for the cosmological constant (Ω_Λ) was almost 10 orders of magnitude smaller than those of either matter (Ω_m) or radiation (Ω_r), so we can neglect its influence in our simple model. The early Universe also contained neutrinos, but their energy density was approximately half that of the photons and they interact so weakly with other components that their influence can also be ignored. We are left with components of radiation and matter, and we will subdivide the latter into baryonic matter and dark matter.

As soon as the primordial density fluctuations appeared, the contents of the Universe began to move under the influence of gravity. The amplitudes of the density perturbations began to increase as dark matter, photons and baryons fell into gravitational potential wells associated with denser-than-average regions of the Universe (often referred to as ‘overdensities’), leaving behind other regions that became correspondingly less dense.

The gravitational influences of matter and radiation do not propagate instantaneously. In fact, the general theory of relativity predicts that information about the distribution of energy density in the universe is transmitted by gravitational waves that propagate at the speed of light. This means that by the epoch of last scattering, the largest structures that would have had time to respond gravitationally to the primordial density fluctuations would have had sizes comparable to the horizon distance at that time:

$$d_{\text{hor}}(t_{\text{ls}}) = a(t_{\text{ls}})c \int_0^{t_{\text{ls}}} \frac{dt}{a(t)} \quad (7.1)$$

Any primordial density perturbations with physical sizes larger than $d_{\text{hor}}(t_{\text{ls}})$ at the epoch of last scattering would not have had time to evolve in this way and should therefore have *retained* their primordial spatial power spectrum.

To calculate a value for the horizon distance in our simplified model Universe by hand, we would need an expression for the scale factor $a(t)$. However, the resulting expression is very complicated and obtaining it

The existence of gravitational waves has now been observationally verified by observatories such as Advanced LIGO and Virgo.

analytically is not straightforward. Instead, try the following Python example, which will take you through the steps needed to compute $d_{\text{hor}}(t_{\text{ls}})$ numerically.

Example 7.1

Using the techniques you learned during the first week-long Python practical activity, write a short program to numerically integrate the Friedmann equation and show that the horizon distance at the epoch of last scattering $d_{\text{hor}}(t_{\text{ls}}) = 0.29 \text{ Mpc}$.

Solution

Please refer to the Example 7.1 Jupyter Notebook in the online module resources to see the solution to this example.

In Example 6.2 you saw that we can use the angular diameter distance $d_A(z_{\text{ls}})$ to relate the size of structures on the surface of last scattering to their apparent angular size when observed from Earth. For structures with a physical size equal to the horizon distance at that time, $d_{\text{hor}}(t_{\text{ls}})$, the corresponding angular size will be

$$\theta_{\text{hor}}(z_{\text{ls}}) = \frac{d_{\text{hor}}(t_{\text{ls}})}{d_A(z_{\text{ls}})} = \frac{0.29 \text{ Mpc}}{12.73 \text{ Mpc}} = 0.023 \text{ rad} \approx 1.3^\circ$$

- To what value of ℓ does an angular size of 1.3° approximately correspond?
- An angular size of 1.3° corresponds to $\ell \approx 180^\circ / 1.3^\circ \approx 140$.

Recall that the normalisation factor of $\ell(\ell + 1)$ in Equation 6.10 is engineered to ensure that primordial density perturbations with a scale-free spatial power spectrum produce a CMB angular power spectrum with an approximately constant value for Δ_T for all values of ℓ . Therefore we would expect the graph in Figure 7.1 to approximate a horizontal line for temperature fluctuations that reflect density perturbations larger than the horizon distance (sometimes referred to as ‘super-horizon’ perturbations), and which correspond to values of ℓ less than ~ 140 .

In reality, the graph does not completely flatten out until $\ell \lesssim 40$, which reflects the fact that $d_{\text{hor}}(t_{\text{ls}})$ is not a hard cut-off and density perturbations larger than the horizon distance would still have evolved slightly prior to the epoch of last scattering. The flat region of the CMB power spectrum below $\ell \approx 40$ is called the **Sachs–Wolfe plateau** in honour of the two scientists, Rainer K. Sachs and Arthur M. Wolfe, whose theoretical predictions helped to explain its existence.

7.1.2 Acoustic oscillations

Primordial perturbations that were smaller than $d_{\text{hor}}(t_{\text{ls}})$ at the epoch of last scattering underwent a very different evolutionary history from those on super-horizon scales. Bear in mind that the horizon distance grows larger as the Universe expands, so density perturbations that were smaller than $d_{\text{hor}}(t_{\text{ls}})$ at the epoch of last scattering may, at earlier times, have been larger than the horizon distance. Smaller perturbations would have been encompassed earlier by the expanding particle horizon and, therefore, would have been evolving for longer at the epoch of last scattering. In fact, the details of a particular perturbation's evolution depend on whether the energy density of the Universe was dominated by matter or radiation when the particle horizon expanded to encompass it.

Exercise 7.1

Consider a universe that contains only matter and radiation. Such a universe is a good model for the real Universe early in its history, when Ω_Λ was negligible. The age $t(a)$ of this universe corresponding to some arbitrarily chosen scale factor, a , can be expressed in terms of the scale factor a_{mr} at the particular instant when the density parameters of matter and radiation were equal:

$$t(a) = \frac{4}{3} \frac{a_{\text{mr}}^2}{H_0 \sqrt{\Omega_{r,0}}} \left[1 - \left(1 - \frac{a}{2a_{\text{mr}}} \right) \left(1 + \frac{a}{a_{\text{mr}}} \right)^{1/2} \right] \quad (7.2)$$

Using this expression and assuming that the epoch of matter–radiation equality occurred at redshift $z_{\text{mr}} \approx 5730$, show that by the epoch of last scattering at $z_{\text{ls}} \approx 1090$ the energy density of this universe had been dark-matter-dominated for just over 94% of its history.

In Chapter 6 you read that before the epoch of last scattering, the opacity of the Universe was very high. This resulted in continuous Thomson scattering between high-energy photons and free electrons, which produced a coupling between the radiation and baryonic matter components that was so strong that they effectively behaved like a single fluid. This medium is sometimes referred to as the **photon–baryon fluid**

In Exercise 7.1 you showed that at the epoch of last scattering, the energy density of the Universe would have been dominated by dark matter for $\sim 90\%$ of its history. The photon–baryon fluid within any density perturbations enveloped by the particle horizon during this time can therefore be considered to have been evolving in a gravitational potential well that was dominated by the influence of dark matter.

Figure 7.2 illustrates how the dark matter drew the photons and baryons with it into the primordial overdensities as they grew. As photons and baryons became concentrated in the evolving overdensities, their temperatures and Thomson scattering rates increased, resulting in

increased internal pressure. Eventually the pressure in the photon–baryon fluid became large enough to counteract the gravitational influence of the dark matter and the fluid began to expand out of the overdense regions at the speed of sound.

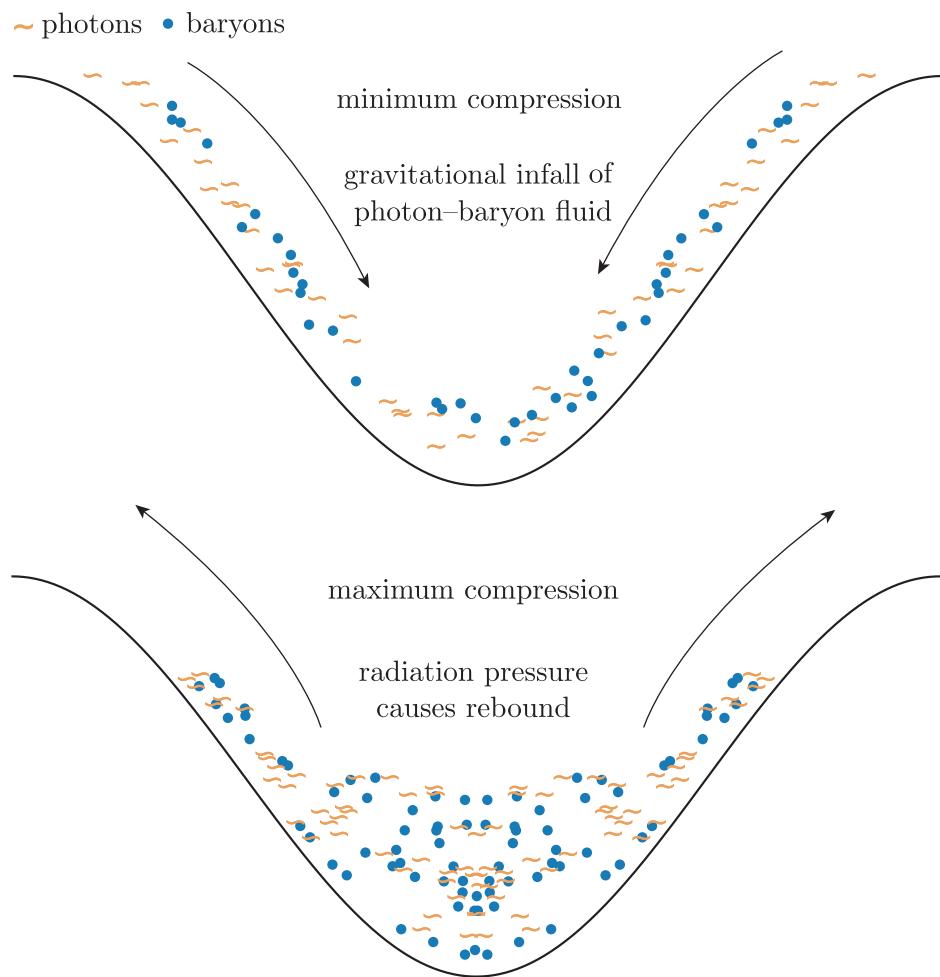


Figure 7.2 Diagram showing how photons and baryons move in the region around a dark-matter overdensity. In the upper panel, a low-pressure photon–baryon fluid begins falling into the gravitational potential well created by the dark matter. As it does so, the internal pressure in the photon–baryon fluid starts to increase. In the lower panel, the radiation pressure has increased so much that it can resist the gravitational influence of the dark matter and the photon–baryon fluid rebounds and expands out of the overdensity at the speed of sound.

- How do you think the expansion of the photon–baryon fluid affected the overall growth of a density perturbation like the one sketched in Figure 7.2 during the matter-dominated era of the Universe’s history?
- The growth of such density perturbations was strongly dominated by the gravitational influence of the dark matter they contained, so the motion of the photons and baryons had almost no influence at all.

As the photon–baryon fluid expanded, its temperature decreased and its internal pressure dropped until it could no longer resist the gravitational influence of the dark matter and the fluid began to fall inward again. This process led to repeated cycles of expansion and collapse that continued until the photons and baryons decoupled, around the epoch of last scattering.

These cycling density and pressure perturbations are the acoustic oscillations that were first introduced in the previous chapter. They are responsible for the peaks that we observe in Figure 7.1. For this reason, the peaks in the CMB power spectrum are often called **acoustic peaks**. You may also see them referred to as *Doppler peaks* but, as we will notice later in the chapter, this is something of a misnomer.

Numerical labels for the acoustic peaks

The acoustic peaks are conventionally labelled using cardinal numbers in order of increasing ℓ . For example, the peak at $\ell \approx 200$ is called the *first* peak, the peak at $\ell \approx 500$ is called the *second* peak, and so on. Later in this chapter we will reference these numerical labels when we discuss the sets of odd- and even-numbered peaks.

In general, the odd-numbered peaks represent oscillations at maximum compression, and the even-numbered peaks correspond to oscillations that are at maximum expansion. By measuring the heights and positions of the different peaks in the CMB power spectrum, cosmologists can derive strong constraints for several cosmological parameters.

A landscape of wells and hills

So far in this section our discussion has focused primarily on regions of the Universe that are denser than average. These overdensities produce gravitational potential wells throughout the Universe. However, when thinking about acoustic oscillations you should keep in mind that for every overdense potential *well* there is a neighbouring potential *hill*, corresponding to a region of the Universe that is less dense than average.

As photons, baryons and dark matter fall into potential wells and increase their density, they are falling from the surrounding potential hills, thereby reducing the density of the latter even further. If we focus on acoustic oscillations with a particular physical scale, the pattern they make throughout the Universe is somewhat like a landscape filled with adjacent *three-dimensional* potential wells and hills, with the photon–baryon fluid sloshing between them.

Figure 7.3 (overleaf) shows a simple representation of this concept in one dimension.

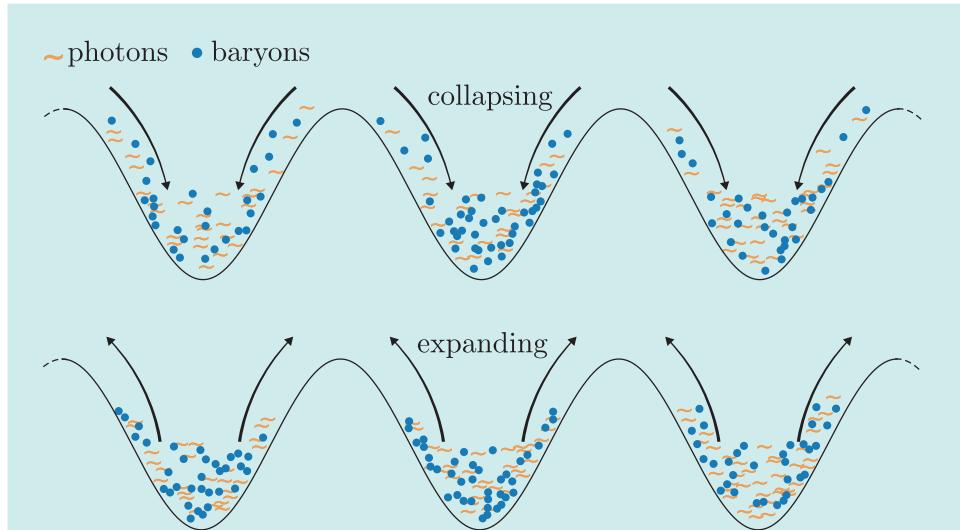


Figure 7.3 Simple one-dimensional illustration showing the landscape of potential wells and hills with the photon–baryon fluid moving within it. In the upper panel the photons and baryons fall from the hills into the wells, and in the lower panel the compressed photon–baryon fluid expands out of the wells and up onto the hills.

However, it is important to remember that this analogy is not perfect. Acoustic oscillations were happening throughout the Universe on *all* physical scales, with smaller oscillations evolving within larger ones. The complex pattern that we observe in the CMB results from the superposition of all these overlapping oscillation patterns.

The physical sizes of the acoustic oscillations were directly proportional to their oscillation frequencies. The acoustic peaks in the CMB angular power spectrum correspond to sets of oscillations that just happened to be at extrema in their oscillation cycles at the epoch of last scattering.

Figure 7.4 shows a schematic of how, during the time before the epoch of last scattering, the baryon–photon fluid pressure varied within acoustic oscillations corresponding to the first three acoustic peaks. The first peak corresponds to oscillations that had only just stopped collapsing for the first time and were about to start their expansion phase, while the second peak was produced by smaller oscillations that were about to recollapse after completing their first expansion phase.

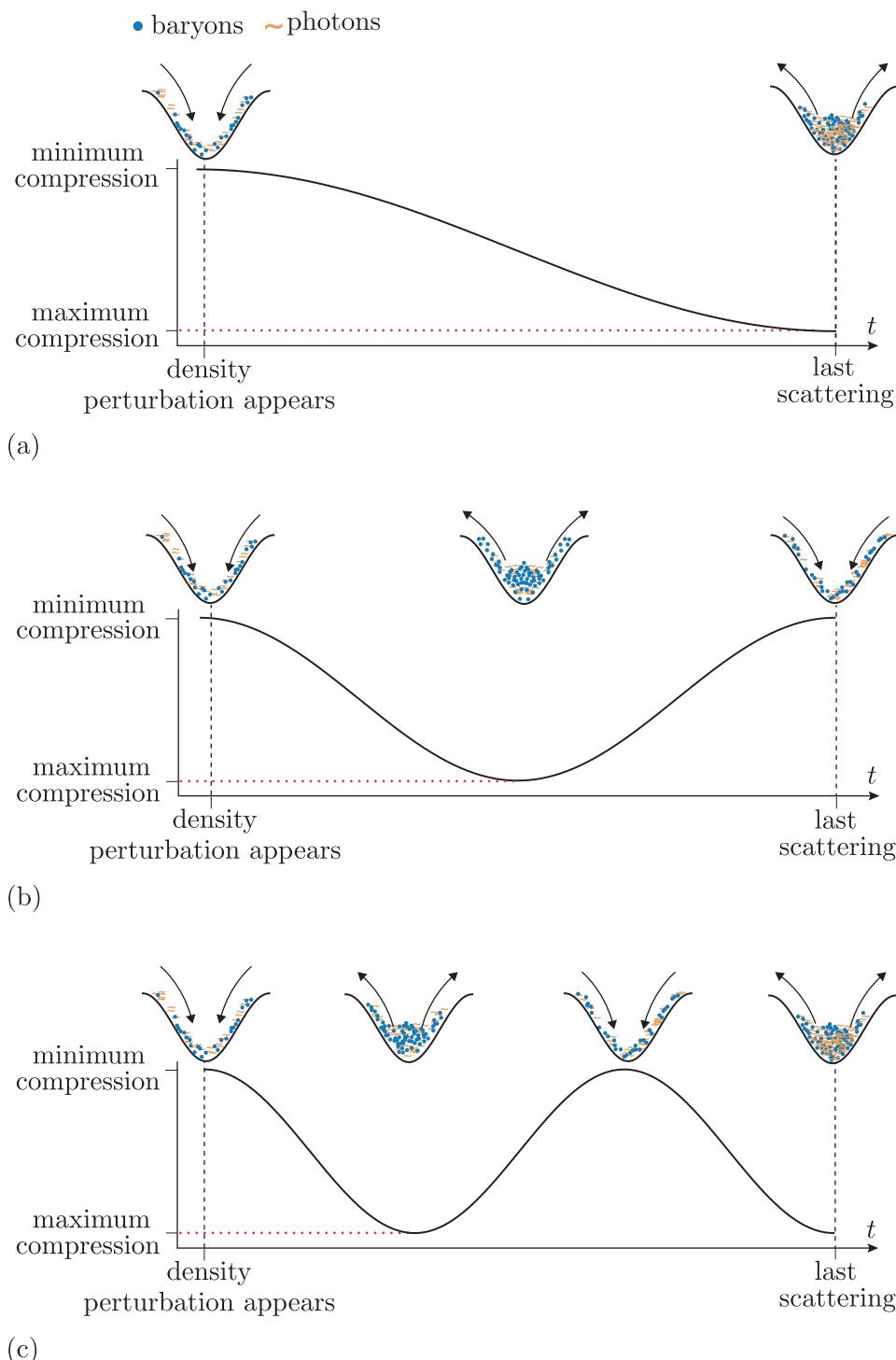


Figure 7.4 Evolution of density over time for three different density perturbations. Panel (a) shows the evolution of density for a perturbation that contributes to the first acoustic peak in the CMB power spectrum; panels (b) and (c) show density perturbations that contribute to the second and third peaks, respectively.

7.1.3 Radiation driving

So far in this section we have focused on the evolution of acoustic oscillations when *dark matter* dominated the energy density of the Universe. Now we briefly discuss the evolution of acoustic oscillations during the short *radiation-dominated* era of the Universe's history and explain how this differs from the dark-matter-dominated case. We will see that radiation dominance can significantly change the way that acoustic oscillations evolve, and we discuss how this impacts the shape of the CMB power spectrum.

For context, bear in mind that the *overall* energy density (sometimes referred to as the background density) decreases as the Universe expands, regardless of whether matter or radiation is dominant. In the matter-dominated case, $\rho \approx \rho_m \propto a^{-3}$, whereas if radiation dominates then the density drops even faster, with $\rho \approx \rho_r \propto a^{-4}$.

Now, recall that acoustic oscillations evolve within *local* regions that are overdense with respect to the universal average. It turns out that the co-evolution between the gravitational potential of these regions and the densities of the fluids they contain depends strongly on whether the Universe is dominated by matter or radiation. A rigorous mathematical description of the co-evolution between gravitational potential and photon density is quite complicated so we will only provide a qualitative description in this book.

For overdensities and their associated potential wells to persist, they must be able to collapse faster than the expanding Universe can reduce their energy density and wipe them out. During the *matter-dominated* era, the gravitational potential of the overdensities was dominated by the energy density of dark matter. These overdensities could persist and even grow because dark matter has no internal pressure to halt its collapse.

In contrast, during the *radiation-dominated* era, the dark matter, baryons and photons were collapsing due to gravitational potentials that were dominated by the energy density of the photons themselves. The increasing internal pressure of the collapsing photons was able to resist the influence of their self-gravity. This slowed the gravitational collapse and allowed expansion of the Universe to start reducing the *local* energy density of photons such that their associated gravitational potential started to decay.

The important result is that during the radiation-dominated era, the potential wells of the overdensities decayed away completely at almost exactly the same time as the photon pressure completely halted their gravitational collapse. The gravitational potential that had compressed the photons and baryons had vanished, but the fluid was still over-pressured relative to its surroundings and started to expand again. Without a gravitational potential to overcome, the fluid was able to keep expanding until its density was substantially lower than it had been before the collapse began (Figure 7.5). The fluid density then continued to oscillate with an amplitude equivalent to half its rebound level.

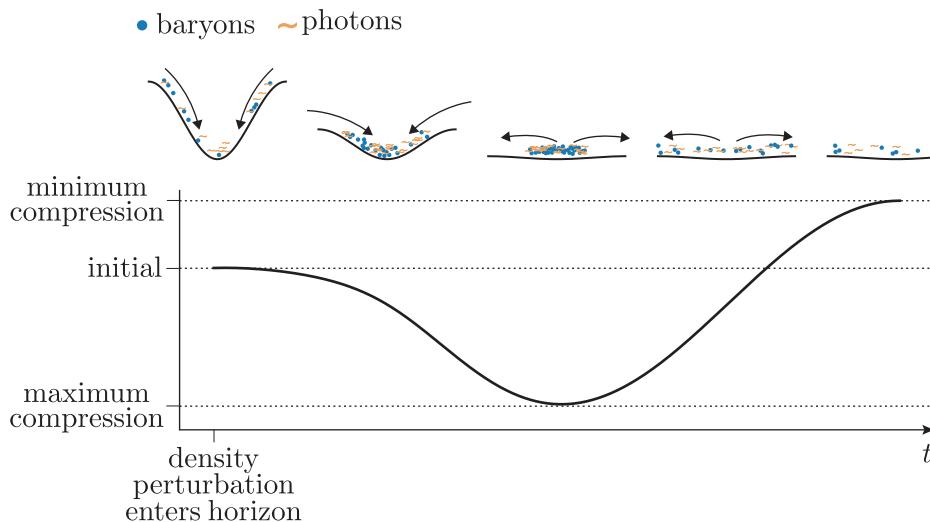


Figure 7.5 Evolution of an acoustic oscillation that becomes encompassed by the particle horizon during the radiation-dominated era of the Universe's history. By the time the photon–baryon fluid reaches maximum compression, the gravitational potential has decayed away and the fluid can rebound to a lower density than it started with.

The overall result is an amplification of acoustic oscillations during the radiation-dominated era, which cosmologists call **radiation driving**. Radiation driving enhances the power spectrum and increases the height of the acoustic peaks for any density perturbations that were smaller than the horizon distance at any point while the Universe's energy density was radiation-dominated.

7.1.4 The origin of the temperature fluctuations

Before we explain how cosmological constraints can be extracted from the CMB power spectrum, we will briefly discuss how the density perturbations that existed at z_{ls} actually produced temperature fluctuations in the CMB.

Consider a population of CMB photons that started their journey to Earth within a dark-matter overdensity at the epoch of last scattering. The gravitational potential of the dark matter redshifts the photons as they propagate out of the overdensity, which makes their spectrum appear cooler. Conversely, CMB photons that were released in underdense regions of the Universe acquired gravitational blueshifts as they fell towards nearby overdensities, making their spectrum appear hotter. This purely gravitational influence on the observed temperatures of CMB photons is called the **Sachs–Wolfe effect** and is the dominant mechanism by which density perturbations that are larger than $d_{\text{hor}}(t_{ls})$ can imprint temperature fluctuations on the CMB.

For density perturbations smaller than $d_{\text{hor}}(t_{\text{ls}})$, the Sachs–Wolfe effect combines with two additional phenomena that also modify the temperature of the CMB photons they produce. The photons and baryons within these *sub-horizon* perturbations were undergoing acoustic oscillations at the epoch of last scattering. Photon fluids that were within the first half of an oscillation cycle had a higher density and were more compressed than the Universal average, while photon fluids in the second half of a cycle had a below-average density.

It is a standard result from thermodynamics that the temperature of a radiation fluid increases as the radiation pressure and energy density increase, such that $T_r \propto P_r^{1/4} \propto \rho_r^{1/4}$. This means that photon fluids that were relatively overdense were hotter than average, while those that were relatively underdense were cooler.

Stefan–Boltzmann law

The energy density ϵ of radiation with temperature T can be calculated using

$$\epsilon = \frac{4\sigma_{\text{SB}}T^4}{c} \quad (7.3)$$

where σ_{SB} is the Stefan–Boltzmann constant.

The *motion* of the photon–baryon fluid as it oscillates also modifies the apparent temperature of the CMB photons by Doppler-shifting their frequencies. If the fluid was receding from us at the epoch of last scattering, then the photons entrained within it will be redshifted and will therefore appear cooler. Conversely, CMB photons that were released from regions of photon–baryon fluid that were moving towards us will be blueshifted and appear hotter. The combination of these two effects means that for photons within dark-matter density perturbations smaller than $d_{\text{hor}}(t_{\text{ls}})$, their observed temperature where they appear on the CMB depends on the phase of their acoustic oscillation at the epoch of last scattering.

In practice, the effect of compression heating the photons dominates over the Doppler shift caused by their motion, which is why the acoustic peaks in Figure 7.1 correspond to maxima and minima in the photon-compression cycle and not to the points where the fluid velocities are at their fastest.

Silk damping

For multipole numbers $\ell \gtrsim 2000$, the heights of the peaks in the CMB power spectrum drop away rapidly. This effect is due to a process called **Silk damping**, after astrophysicist Joseph Silk.

In Section 6.1 we mentioned that the Universe did not become transparent instantaneously and so the surface of last scattering is really more like a thin shell. As the opacity of the Universe decreased, the photon–baryon fluid began to decouple and photons were able to travel larger distances before they were scattered by electrons. This allowed them to exchange energy with more distant photons and baryons, which in turn reduced the difference in temperature between nearby density perturbations.

The effect on the CMB is to smear out temperature fluctuations that correspond to physical scales smaller than the thickness of the last-scattering ‘shell’. This smearing reduces the variance of the CMB on these scales and suppresses the peaks in the high- ℓ region of the power spectrum.

By measuring the multipole number at which Silk damping manifests, cosmologists can determine how thick the last scattering shell is and, therefore, how long it took for the Universe to become transparent.

7.2 Cosmological parameter measurements

In this section you will finally learn how the observed pattern of CMB temperature fluctuations can provide strong constraints on the cosmological parameters of our Universe.

You already read in Section 6.3.2 that models of the Universe can be used to predict the expected values for different statistical properties of the CMB, including its angular power spectrum C_ℓ . In this section we will spend some time discussing a series of figures that show the effect of varying particular cosmological parameters (Ω_k , $\Omega_{b,0}$, $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$) on the CMB angular power spectrum. By comparing the measured power spectrum with theoretical predictions from models that assume different sets of cosmological parameters, cosmologists can identify the sets of models and model parameters that best fit the observed data.

7.2.1 The acoustic scale

The acoustic oscillations that were described in Section 7.1.2 produce a pattern of fluctuating density and pressure throughout the Universe. Fluctuations in density and pressure are just sound waves, and the pattern of acoustic oscillations can be treated as a superposition of standing sound waves with different wavelengths that correspond roughly to the spacing between neighbouring dark-matter overdensities.

Now, consider a particular acoustic oscillation that contributes to the first acoustic peak in the CMB angular power spectrum. If we could use a

cosmological model to predict the physical size of this oscillation, then we can use Equation 7.1 to compute the expected angular size of the CMB temperature fluctuation it produces. By comparing our predicted angular size with the observed position of the first acoustic peak in the CMB angular power spectrum, we can find a set of parameters for our cosmological model that match those of the real Universe. The calculation required to compute this oscillation's physical size lies outside the scope of this module, but the solution is that it is roughly equal to the **sound horizon** at the epoch of last scattering, which we will denote using the symbol d_s and which can be calculated using:

$$d_s(z_{ls}) = a(z_{ls}) \int_0^{t_{ls}} \frac{c_s}{a(t)} dt \quad (7.4)$$

The sound horizon represents the maximum proper distance that a sound wave could have propagated between the big bang and the epoch of last scattering. The speed of sound c_s during this early phase of the Universe's evolution can be calculated using assumed model values for the baryon density parameter Ω_b and the radiation density parameter Ω_r at that time:

$$c_s = c \left(3 + \frac{9}{4} \frac{\Omega_b}{\Omega_r} \right)^{-1/2} \quad (7.5)$$

The second, third and fourth acoustic peaks correspond to oscillations that have physical sizes one-half, one-quarter and one-eighth of $d_s(z_{ls})$, respectively. To get some physical intuition for this result, recall from Section 7.1.2 that oscillations contributing to the first acoustic peak were just about to rebound after their first collapse phase, which began when the primordial density perturbations first appeared in the Universe. Reaching a state of maximum compression implies that a propagating sound wave had time to modify the fluid pressure throughout the whole oscillation, which means that its maximum possible size is roughly equal to the sound horizon.

We can use the angular diameter distance to the surface of last scattering $d_A(z_{ls})$ to calculate a quantity called the **acoustic scale**, which we will denote θ_s and is defined as the observable angular scale of temperature fluctuations contributing to the first acoustic peak in Figure 7.1.

The acoustic scale

$$\theta_s \approx \frac{d_s(z_{ls})}{d_A(z_{ls})} \quad (7.6)$$

The acoustic scale is one of the most important directly observable properties of the CMB and its value can be used to help constrain several of the cosmological parameters.

Example 7.2

Show that Ω_b remained less than Ω_r until after the epoch of last scattering.

Solution

To solve this problem we first need to compute the redshift z_{rb} when the energy densities of baryons and radiation were equal. At all higher redshifts, corresponding to smaller scale factors, the radiation energy density would have dominated because $\rho_r \propto a^{-4}$, whereas the baryon density evolves like matter, that is $\rho_b \propto a^{-3}$. We want to show that $z_{rb} < z_{ls}$.

To find z_{rb} , we start by computing the corresponding scale factor a_{rb} . Because we know that the densities of baryons and radiation are equal when $a = a_{rb}$, we can write:

$$\rho_r(a_{rb}) = \rho_b(a_{rb})$$

Using the density power laws for radiation and matter (baryonic) gives

$$\frac{\rho_{r,0}}{a_{rb}^4} = \frac{\rho_{b,0}}{a_{rb}^3}$$

and we then divide both sides by the present-day critical density $\rho_{c,0}$ (defined in Equation 4.27):

$$\frac{\rho_{r,0}}{a_{rb}^4 \rho_{c,0}} = \frac{\rho_{b,0}}{a_{rb}^3 \rho_{c,0}}$$

Identifying the ratio of densities on each side as the present-day density parameters and cancelling scale factors gives

$$\frac{\Omega_{r,0}}{a_{rb}} = \Omega_{b,0}$$

and so

$$a_{rb} = \frac{\Omega_{r,0}}{\Omega_{b,0}}$$

Now we just need to compute z_{rb} . Using the values listed in the table of constants:

$$\begin{aligned} z_{rb} &= \frac{1}{a_{rb}} - 1 = \frac{\Omega_{b,0}}{\Omega_{r,0}} - 1 \\ &= \frac{0.0490}{5.4 \times 10^{-5}} - 1 \\ &\approx 910 \end{aligned}$$

As required, this value is less than $z_{ls} \approx 1090$ so we have shown that Ω_b remained less than Ω_r until after the epoch of last scattering.

Exercise 7.2

In Example 7.2 you saw that Ω_b remained less than Ω_r until after the epoch of last scattering. This result means that you can assume baryons have a negligible influence on the properties of the photon–baryon fluid, i.e. that $\Omega_r \gg \Omega_b$.

- (a) Hence, show that the approximation

$$d_s(z_{ls}) \approx \frac{d_{\text{hor}}(z_{ls})}{\sqrt{3}}$$

where $d_{\text{hor}}(z_{ls})$ is the horizon distance at t_{ls} (when photons with observed redshift z_{ls} were emitted), is a good one.

- (b) Now use this approximation and the fact that $d_A(z_{ls}) = 12.73 \text{ Mpc}$ to estimate the multipole ℓ_s that corresponds to the acoustic scale in the CMB power spectrum; you may assume the result from Example 7.1 that $d_{\text{hor}}(z_{ls}) = 0.29 \text{ Mpc}$.

Your answer should be very close to the *observed* location of the first peak in Figure 7.1.

7.2.2 Measuring Ω_k

We end this chapter with some examples of cosmological parameters that can be constrained using the properties of the CMB angular power spectrum.

Let's consider the measured location of the first acoustic peak, which corresponds to the acoustic scale and can be used to constrain the *curvature* of the Universe. As discussed in Chapter 5, curvature affects the measured angular sizes for objects of a fixed physical size. This includes the CMB acoustic peaks, whose physical scale is fixed by the physics discussed in the previous section. The predicted effect of spatial curvature on the CMB power spectrum is shown in Figure 7.6. Increasing the value of the density parameter for curvature, Ω_k , makes temperature fluctuations seem smaller and moves all of the acoustic peaks to higher values of ℓ .

- Bearing in mind the curves shown in Figure 7.6, is the CMB power spectrum measured by *Planck* compatible with $\Omega_k < -0.2$?
- No. Notwithstanding the fact that the predicted peak positions do not match their observed counterparts, values of $\Omega_k < -0.2$ predict a very large rise at low values of ℓ that is strongly disfavoured by the observed data, which are plotted in Figure 7.1.

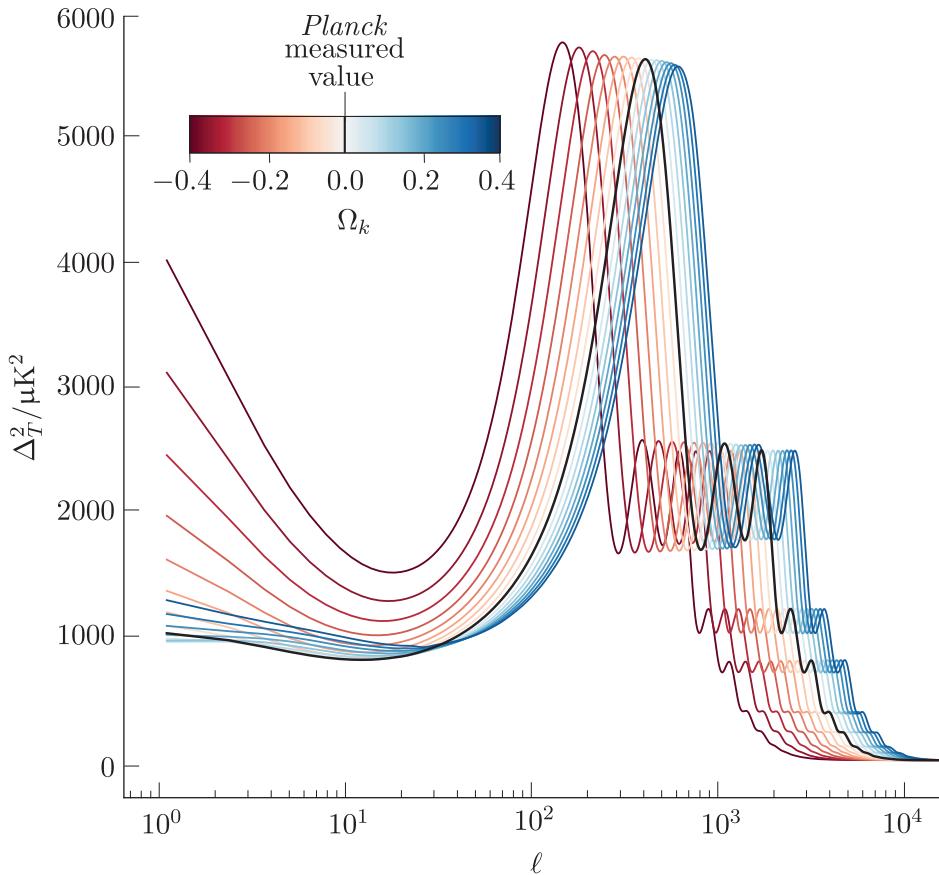


Figure 7.6 Effect of varying Ω_k on the CMB power spectrum. The value of Ω_k derived from *Planck*'s measurements is shown as a solid black line; all other cosmological parameters are fixed to the values listed in the table of constants.

The value for $d_A(z_{ls}) = 12.73 \text{ Mpc}$ that you assumed in Exercise 7.2 assumes a flat universe with $\Omega_k = 0$. The fact that the position you estimated for the first acoustic peak agrees so well with observation is compelling evidence that the real Universe is indeed spatially flat.

The precise positions of the acoustic peaks in Figure 7.1 also depend somewhat on the relative densities of photons and baryons in the Universe at and before the epoch of last scattering. This is because both Ω_b and Ω_r appear in the definition of the speed of sound in Equation 7.5. Unless we can find a way to independently constrain Ω_b/Ω_r , this additional dependency will limit how tightly we can constrain Ω_k . There might be several different sets of values for Ω_b , Ω_r and Ω_k that predict very similar locations for the acoustic peaks and we would have no way of knowing which set was true! Fortunately, the CMB power spectrum *itself* provides another, largely independent way to determine Ω_b , by comparing the *amplitudes* of the different acoustic peaks. We discuss this further in the next section.

7.2.3 Measuring $\Omega_{b,0}$

In this section we will discuss how observations of the CMB power spectrum can also be used to constrain the *baryonic* matter density parameter $\Omega_{b,0}$.

Figure 7.7 shows how changing the value of $\Omega_{b,0}$, while keeping the values of $\Omega_{m,0}$, $\Omega_{\Lambda,0}$ and H_0 fixed to the values in the table of constants, affects the *relative* heights of the acoustic peaks. A higher baryon fraction enhances the first and third acoustic peaks, but reduces the height of the second.

Constraints on the density parameters that are derived from CMB data alone are often presented as multiples of h^2 where h represents the Hubble constant, H_0 , in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

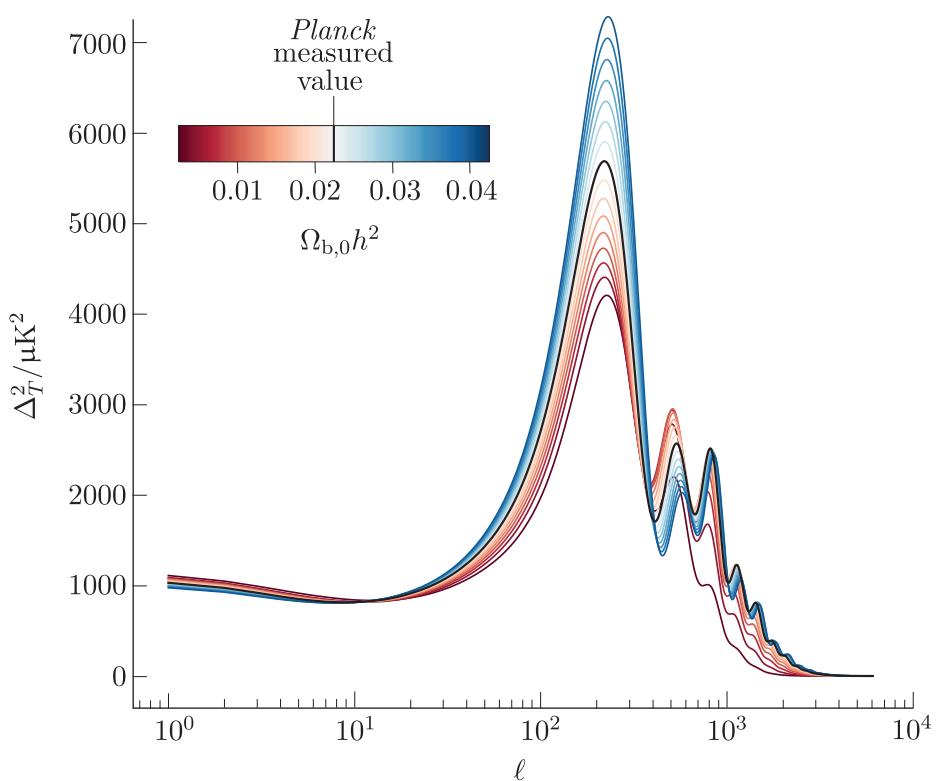


Figure 7.7 The effect of varying $\Omega_{b,0}$ on the CMB power spectrum. The value of $\Omega_{b,0}$ derived from *Planck*'s measurements is shown as a solid black line; all other cosmological parameters are fixed to the values listed in the table of constants.

To understand these effects we need to recognise that adding baryons to the photon–baryon fluid increases its mass but does not change its pressure. This is because baryons are matter and have sufficiently low density that their equation of state parameter $w_b = 0$ (see Chapter 4). A useful analogy for the effect of increasing Ω_b is illustrated in Figure 7.8. The oscillating photon–baryon fluid behaves in a similar way to a mass attached to a spring, oscillating in the vertical direction under gravity. The spring represents the radiation pressure counteracting the gravitational influence of the dark-matter overdensity.

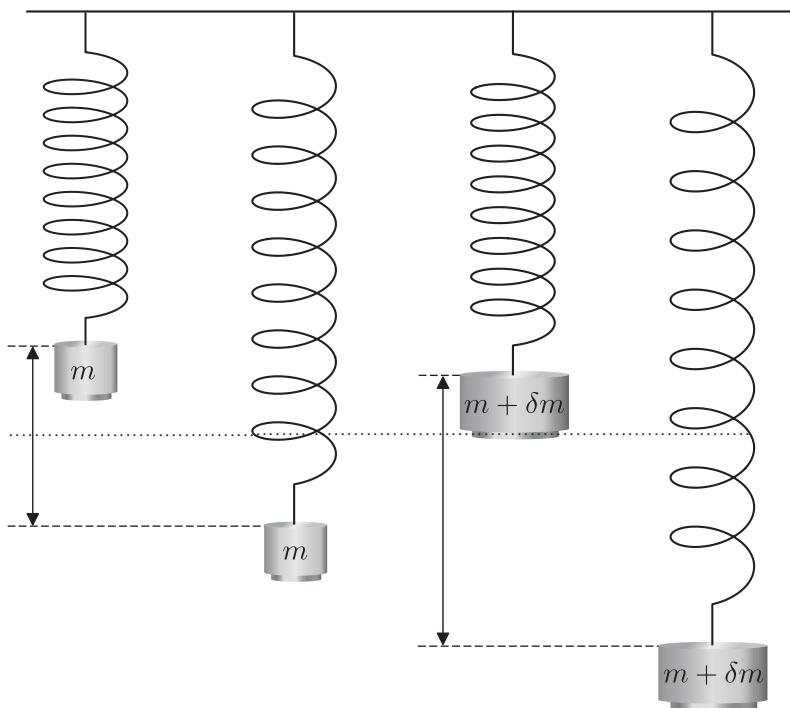


Figure 7.8 Masses oscillating on a spring. Increasing the mass while leaving the spring unchanged is analogous to increasing the density of baryons in an acoustic oscillation. The effect is to increase the amplitude of the oscillations (shown by the vertical arrows) and also to lower their midpoint. The dotted line represents the oscillation midpoint of the smaller mass and is analogous to the mean density of the Universe.

Increasing the size of the mass while leaving the spring unchanged has two effects on the oscillations it undergoes. First, the mass falls further before the spring can arrest its motion, so the amplitude of the oscillations increases. In a similar way, a more massive photon–baryon fluid will sink deeper into the gravitational potential of a dark-matter overdensity before its internal pressure can halt its collapse. This means that when the fluid reaches its maximum compression, it will be denser and the photons within it will have a higher temperature.

- Does this mean that we should expect to see more CMB hot spots than cold spots in a Universe that contains more baryons?
- No. Remember that for every potential well that fills up with photons and baryons there is a corresponding potential hill that drains away. If the fluid in the wells becomes more compressed, then the fluid on the hills must become even more rarefied.

Most importantly, the *difference* between the temperature of the oscillations that are at maximum compression and the average temperature of the Universe is larger if the Universe contains more baryons. This increases the contribution to the variance of the CMB of oscillations that had reached maximum compression at the epoch of last scattering, which enhances the angular power spectrum on angular scales that correspond to the odd-numbered acoustic peaks.

The second impact of increasing the mass attached to the spring is that the midpoint of the oscillation, where the speed of the mass is greatest, moves lower. A similar effect modifies the acoustic oscillations. During the expansion phase of the acoustic oscillations the pressure from photons in the fluid must now overcome the added inertia from the extra baryons. This means that the fluid will not rebound as far, and oscillations that had reached their maximum rarefaction at the epoch of last scattering will have photon temperatures that are closer to the Universal average. The result of adding more baryons to the Universe is that the even-numbered peaks in the CMB angular power spectrum are suppressed.

Figure 7.9 shows a comparison of how the density of an acoustic oscillation within a potential well evolves for fluids that contain baryons and for those that do not.

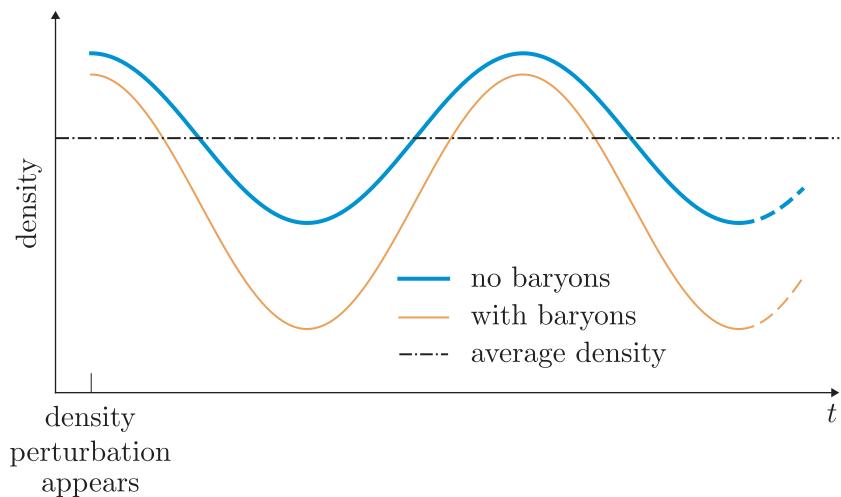


Figure 7.9 The evolution of density in a potential *well* before t_{ls} for fluids that contain baryons (orange) and those that do not (blue). Adding baryons increases the amplitude of the oscillations and lowers the oscillation midpoint with respect to the average density of the Universe (dot-dashed line).

- If adding baryons increases the maximum pressure and temperature of the photon–baryon fluid in the potential wells, why don't the average temperature and density of the Universe also go up?
- Again, we need to remember that every potential well has a corresponding potential hill. If baryons and photons sink deeper into a potential well, they fall further from the potential hill, leaving it more rarefied and therefore cooler. The temperature and density changes in the wells and on the hills balance each other out, so the *average* temperature and density stay the same.

By measuring the heights *and* positions of the acoustic peaks observed by the *Planck* satellite, cosmologists are able to very precisely constrain the product $\Omega_{b,0}h^2 = 0.022\,42 \pm 0.000\,14$, where h represents the Hubble constant, H_0 , in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

7.2.4 Measuring $\Omega_{\text{m},0}$

In this section we will discuss how observations of the CMB power spectrum can also be used to constrain the *overall* matter density parameter $\Omega_{\text{m},0}$.

Figure 7.10 shows the effect that changing $\Omega_{\text{m},0}$ has on the CMB power spectrum. Increasing $\Omega_{\text{m},0}$ decreases the amplitude of the acoustic peaks, while decreasing $\Omega_{\text{m},0}$ has the opposite effect. This variation is caused by the radiation-driving phenomenon that you read about in Section 7.1.3.

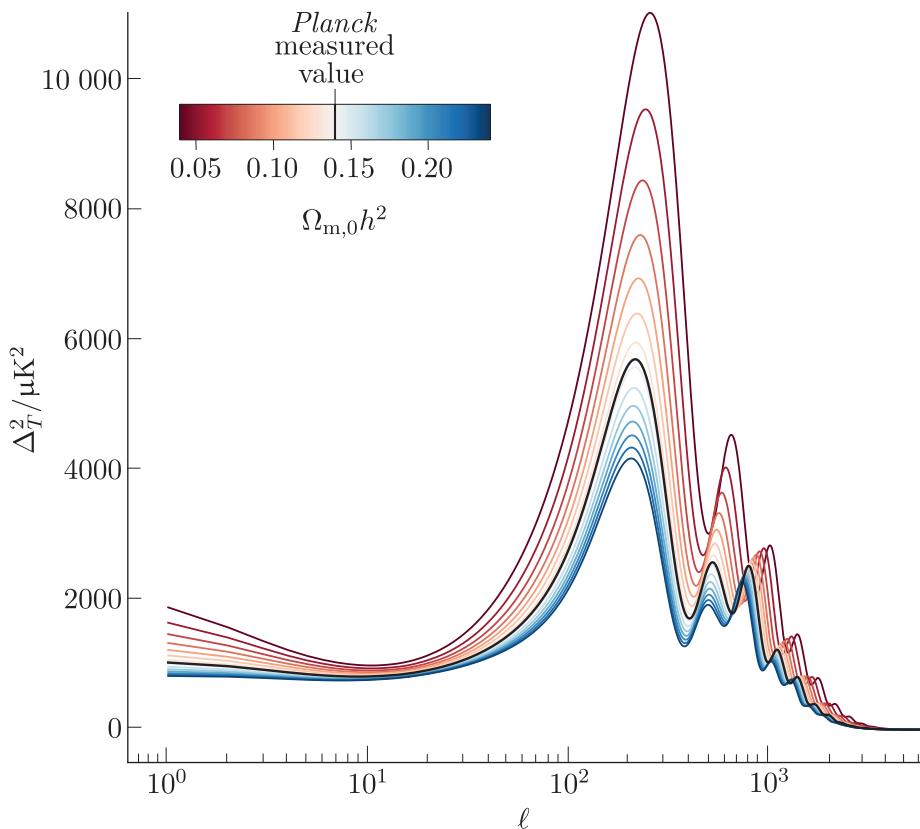


Figure 7.10 Effect of varying $\Omega_{\text{m},0}$ on the CMB power spectrum. The value of $\Omega_{\text{m},0}$ derived from *Planck*'s measurements is shown as a solid black line; all other cosmological parameters are fixed to the values listed in the table of constants.

- What percentage of *matter* in the Universe is baryonic? What other type of matter constitutes the remainder?
- You may recall from Chapter 1 that about 17% of matter in the Universe is baryonic. The rest is non-baryonic dark matter.

The primary effect of increasing $\Omega_{\text{m},0}$ is to increase the amount of dark matter within overdensities. During the radiation-dominated era of the Universe's history, this helped to stabilise them and prevent potential wells from decaying away completely. The residual gravitational potentials reduce the rebound amplitude of the acoustic oscillations, thereby reducing

the heights of the acoustic peaks. Decreasing $\Omega_{m,0}$ has the opposite effect: potential wells decay away more quickly, leaving them shallower and allowing acoustic oscillations to expand further and reach lower densities when they rebound.

- Would you expect radiation driving to amplify the CMB power spectrum on all angular scales?
- No. Only oscillations that were smaller than the horizon distance during the radiation-dominated era of the Universe's history would be amplified. This means that oscillations larger than the angular scale corresponding to the horizon distance at the epoch of matter–radiation equality would not be amplified.

Varying the overall matter density also shifts the *positions* of the acoustic peaks by changing the size of the sound horizon at the epoch of last scattering. By accounting for the observed positions and amplitudes of the acoustic peaks, the CMB power spectrum can be used to constrain $\Omega_{m,0}h^2$ to within 2%.

7.2.5 Measuring $\Omega_{\Lambda,0}$

Finally in this section we consider the energy density of the cosmological constant. We will find that CMB data *on their own* can only provide very weak constraints on $\Omega_{\Lambda,0}$ and we will discuss why this is the case.

Figure 7.11 shows that large changes in $\Omega_{\Lambda,0}$ produce minimal variations in either the positions *or* the amplitudes of the acoustic peaks in the CMB power spectrum.

The main reason that the value of $\Omega_{\Lambda,0}$ has so little influence on the CMB power spectrum is that the energy density of the cosmological constant was completely negligible compared to those of matter and radiation during the time before t_{ls} , when the acoustic oscillations were evolving. The following example demonstrates the huge dominance of matter and radiation over Λ prior to the epoch of last scattering.

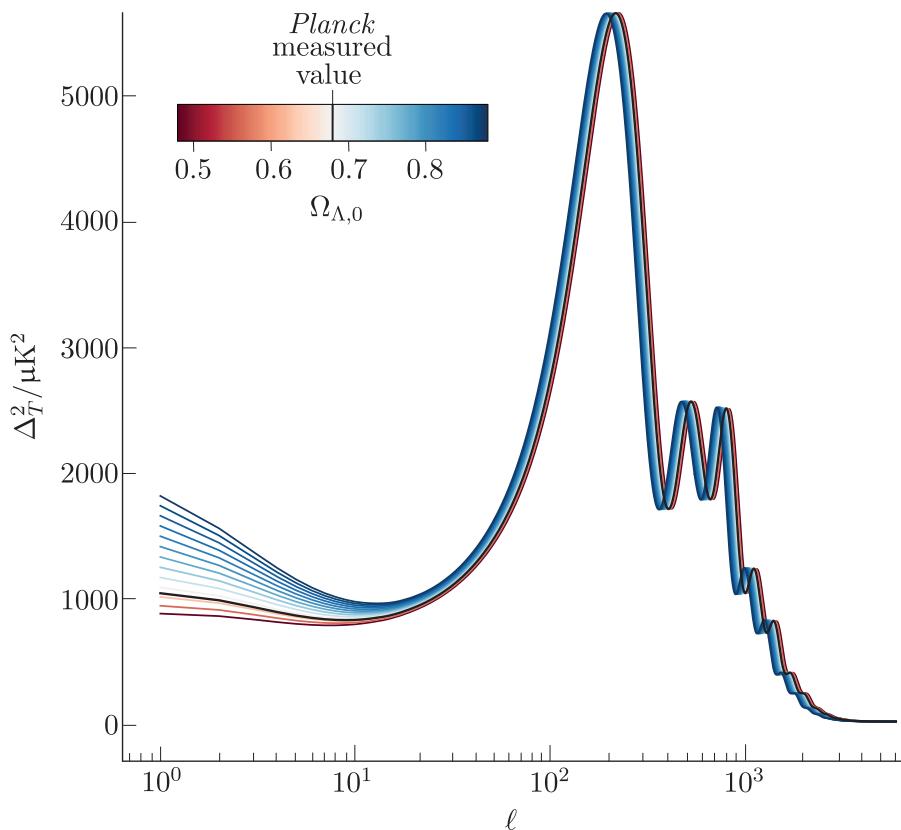


Figure 7.11 Effect of varying $\Omega_{\Lambda,0}$ on the CMB power spectrum. The value of $\Omega_{\Lambda,0}$ derived from *Planck*'s measurements is shown as a solid black line; all other cosmological parameters are fixed to the values listed in the table of constants.

Example 7.3

- Using the values listed in the table of constants, compute the values of Ω_m/Ω_r and Ω_m/Ω_Λ at the epoch of last scattering, assuming $z_{ls} = 1090$.
- How would these ratios evolve for even higher redshifts?

Solution

- To solve this problem we can relate the ratios of the present-day density parameters to their values at z_{ls} . We first need to recall from Chapter 4 how the density of each component evolves as the Universe expands:

$$\rho_m = \rho_{m,0}a^{-3} = \rho_{m,0}(1+z)^3$$

$$\rho_r = \rho_{r,0}a^{-4} = \rho_{r,0}(1+z)^4$$

$$\rho_\Lambda = \rho_{\Lambda,0}$$

Now we can compute the required ratios:

$$\frac{\Omega_m(z_{ls})}{\Omega_r(z_{ls})} = \frac{\rho_m(z_{ls})/\rho_c(z_{ls})}{\rho_r(z_{ls})/\rho_c(z_{ls})} = \frac{1}{1+z_{ls}} \frac{\rho_{m,0}}{\rho_{r,0}} = \frac{1}{1+z_{ls}} \frac{\Omega_{m,0}}{\Omega_{r,0}}$$

and

$$\frac{\Omega_m(z_{ls})}{\Omega_\Lambda(z_{ls})} = \frac{\rho_m(z_{ls})/\rho_c(z_{ls})}{\rho_\Lambda(z_{ls})/\rho_c(z_{ls})} = (1+z_{ls})^3 \frac{\rho_{m,0}}{\rho_{\Lambda,0}} = (1+z_{ls})^3 \frac{\Omega_{m,0}}{\Omega_{\Lambda,0}}$$

The final steps in each of the previous two expressions replace the ratio of present-day densities with the (equivalent) ratio of present-day density parameters. Using the values from the table of constants, we find

$$\frac{\Omega_m(z_{ls})}{\Omega_r(z_{ls})} = \frac{1}{1+z_{ls}} \frac{\Omega_{m,0}}{\Omega_{r,0}} = \frac{1}{1091} \frac{0.3097}{5.4 \times 10^{-5}} = 5.3$$

and

$$\frac{\Omega_m(z_{ls})}{\Omega_\Lambda(z_{ls})} = (1+z_{ls})^3 \frac{\Omega_{m,0}}{\Omega_{\Lambda,0}} = (1091)^3 \frac{0.3097}{0.6888} = 5.84 \times 10^8$$

- (b) The density of matter grows more slowly with redshift than the density of radiation does. Therefore at higher redshifts the value of Ω_m/Ω_r would decrease and fall below 1 at z_{mr} , when the energy densities of matter and radiation become equal. Conversely, Ω_m/Ω_Λ would continue to increase all the way back to the big bang.

Bearing in mind the results that are derived in Example 7.3, the Friedmann equation tells us that the value of $\Omega_{\Lambda,0}$ has almost no influence on the expansion of the Universe during the time before the epoch of last scattering. As a consequence, the cosmological constant barely affects physical scales that impact the CMB power spectrum (e.g. the sizes of the sound horizon and the particle horizon) and it does not enhance or diminish the radiation-driving effect.

In addition to being very small, the effect of varying $\Omega_{\Lambda,0}$ on the CMB power spectrum is almost identical to the effect of varying Ω_k . Both parameters shift the locations of the peaks without changing their relative spacing or substantially changing their amplitudes. A very small amount of curvature can replicate large changes in the energy density of Λ , which means that it is impossible to *simultaneously* derive precise constraints on $\Omega_{\Lambda,0}$ and Ω_k using CMB data *alone*.

7.2.6 Cosmological parameter degeneracies and combined constraints

In the previous section you saw how observations of the CMB can be used to provide strong constraints on many of the cosmological parameters. However, you also saw that models with different sets of input parameters can predict CMB power spectra with very similar observable characteristics, and that varying different parameters can sometimes

modify those predictions in very similar ways. Such correlations between the effects of different parameters on the output of a model are called **parameter degeneracies**, and the parameters that produce the correlation are said to be **degenerate**.

Degeneracies between the cosmological parameters limit our ability to precisely constrain their values using observational data, no matter how exquisite those data are. For example, consider the orange region in Figure 7.12a, which highlights the ranges of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that are consistent with CMB observations made using the *WMAP* telescope. The overall extent of this region in the horizontal and vertical directions indicates the precision of the parameter constraints that the *WMAP* data provide; a smaller extent in either direction implies a more precise constraint on the corresponding cosmological parameter.

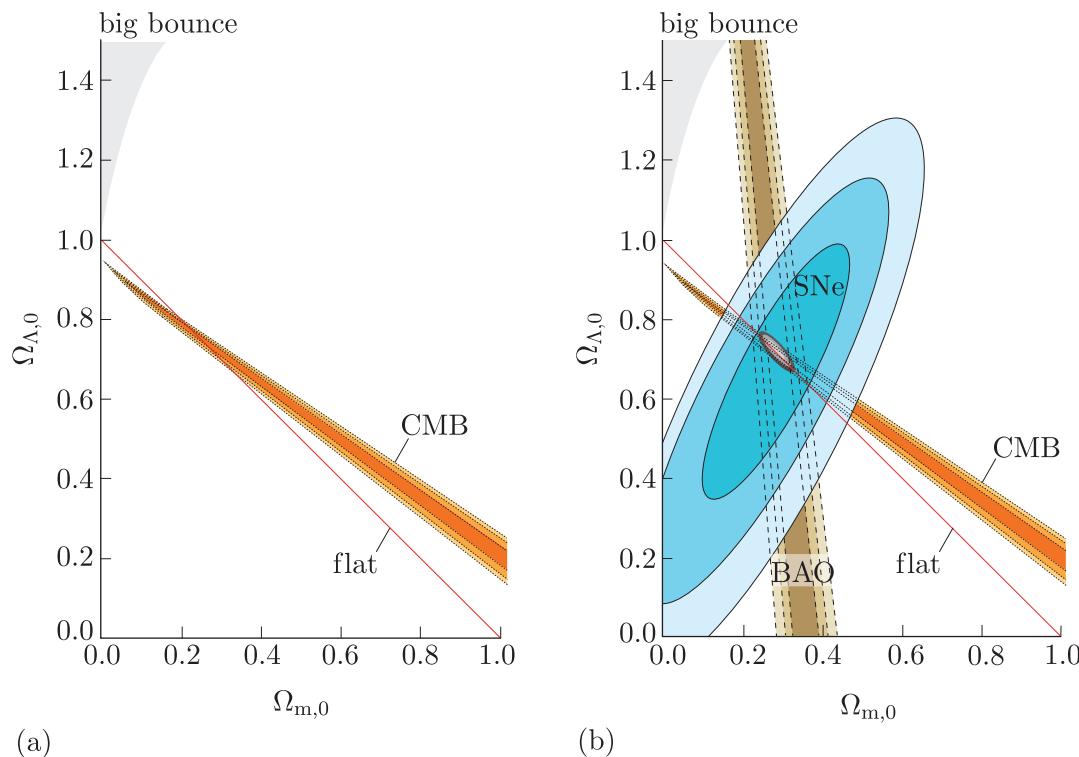


Figure 7.12 (a) Constraints on $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ derived using observations of the CMB power spectrum on its own (orange contours). (b) Constraints on $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ derived using observations of Type Ia supernovae (SNe; blue contours), baryon acoustic oscillations (BAO; brown contours) *and* the CMB (orange contours). The small grey oval in (b) shows the region of joint agreement between all three observational measurements. In both panels, the different contour levels indicate regions of decreasing uncertainty, from 3σ (lightest tone) to 2σ and 1σ (darkest).

The orange region is very narrow because for any *specific* value of $\Omega_{m,0}$ (or $\Omega_{\Lambda,0}$), there is a very small range of $\Omega_{\Lambda,0}$ (or $\Omega_{m,0}$) values that are permitted by the *WMAP* data. However, the region is also very extended and spans a large range of parameter values because the values of $\Omega_{m,0}$ and

$\Omega_{\Lambda,0}$ are both *degenerate* with several other cosmological parameters in terms of their effect on the predicted CMB power spectra, primarily Ω_k , $\Omega_{b,0}$ and H_0 . Varying the value of $\Omega_{m,0}$ does change the predicted shape of the power spectrum, but this change can be almost completely undone by choosing appropriate values of $\Omega_{\Lambda,0}$, Ω_k , $\Omega_{b,0}$ and H_0 .

- Look again at Figures 7.6, 7.10 and 7.11. What property of the CMB power spectrum is similarly affected by variations in Ω_k , $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that might therefore be contributing to the degeneracy among these parameters?
- The *positions* of the acoustic peaks are similarly affected; increasing either $\Omega_{m,0}$ or $\Omega_{\Lambda,0}$ shifts the peaks to lower multipoles, but this shift can be undone by increasing Ω_k .

There are several techniques that can be used to reduce or circumvent the degeneracies between different cosmological parameters. Such techniques are said to *break* the degeneracies and they allow cosmologists to constrain the cosmological parameters much more precisely.

One commonly used way to break degeneracies in a model with a large number of parameters is to fix one or more of those parameters to particular values. For example, we have already seen in Section 7.2.5 that Ω_k and $\Omega_{\Lambda,0}$ are degenerate parameters that shift the acoustic peaks in a similar way. However, if we *assume* a flat Universe with $\Omega_k = 0$, then we can break this degeneracy and derive a good constraint on $\Omega_{\Lambda,0}$. The solid red diagonal line (labelled ‘flat’) in the panels of Figure 7.12 identifies the values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ that are consistent with a spatially flat Universe. It represents the Friedmann equation (Equation 4.40) for a Universe with zero curvature and negligible $\Omega_{r,0}$:*

$$1 = \Omega_{m,0} + \Omega_{\Lambda,0} + \Omega_{r,0} \approx \Omega_{m,0} + \Omega_{\Lambda,0} \quad (7.7)$$

The assumption of a flat Universe reduces the set of permissible parameter values to the small range in which the red line intersects the orange contours. If Ω_k is assumed to be zero, then the *WMAP* satellite data constrain $\Omega_{\Lambda,0}$ to be 0.679 ± 0.013 .

Even if cosmological parameters are degenerate with respect to their effect on the CMB, it is often possible to break those degeneracies using auxiliary constraints that are derived from observations of phenomena other than the CMB. For example, in Chapter 5 you learned that observations of Type Ia supernovae and measurements of baryon acoustic oscillations (BAOs) can also be used to place constraints on the values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$. Figure 7.12b shows these constraints, together with those that can be derived using CMB observations.

The different observational constraints exclude different possible values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$. The resultant combined constraints are indicated by the small grey area in Figure 7.12b, which lies at the centre of the region

* $\Omega_{r,0}$ is much smaller than either $\Omega_{m,0}$ or $\Omega_{\Lambda,0}$ so it can usually be neglected in practice. Otherwise, its value is strongly constrained by the observed average CMB temperature.

where the three independent observational constraints overlap. Notice that this grey area coincides closely with the solid red line even though no prior assumptions have been made about the value of Ω_k . The combined data from the CMB, supernovae and BAOs provide strong *observational* evidence that the Universe is indeed flat, and they constrain Ω_k extremely precisely, to a value of 0.0007 ± 0.0019 .

In this chapter you have seen that the CMB encodes a wealth of information about the cosmological parameters and the physics of the early Universe. However, it is only by *combining* the CMB observations with complementary observational datasets that the exquisitely precise constraints listed in the table of constants can be obtained. It is often said that we are now living in an age of ‘precision cosmology’. Access to such precise measurements allows us to thoroughly test our theoretical models of the Universe and search for evidence of new, unexpected physical processes and phenomena.

7.3 Summary of Chapter 7

- The positions and amplitudes of the **acoustic peaks** in the CMB angular power spectrum reflect the physical sizes and amplitudes of acoustic oscillations at the epoch of last scattering.
- The physical sizes and amplitudes of the acoustic oscillations depend on the speed of sound and physical conditions in the early Universe, which in turn depend on the values of the cosmological parameters. This means that the observed positions and amplitudes of the acoustic peaks in the CMB angular power spectrum depend indirectly, but predictably, on the values of the cosmological parameters.
- Precise measurement and careful analysis of the CMB angular power spectrum can be used to constrain the values of several cosmological parameters, including $\Omega_{b,0}$, $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$.
- Some cosmological parameters affect the predicted CMB angular power spectrum in very similar ways. These groups of parameters are said to be **degenerate**. In particular, $\Omega_{\Lambda,0}$, $\Omega_{m,0}$ and Ω_k are degenerate with respect to their effect on the predicted CMB power spectrum. This degeneracy weakens the constraints on these parameters that can be *simultaneously* derived from CMB observations *alone*.
- **Parameter degeneracies** can be broken by assuming specific values for certain cosmological parameters. For example, assuming that $\Omega_k = 0$ breaks the degeneracy between $\Omega_{\Lambda,0}$ and $\Omega_{m,0}$.
- The degeneracy between $\Omega_{\Lambda,0}$, $\Omega_{m,0}$ and Ω_k can also be broken via complementary constraints derived from observations of Type Ia supernovae and/or baryon acoustic oscillations.

Chapter 8 Physics of the early Universe

The previous three chapters took an in-depth look at how astronomical observations enable us to measure cosmological parameters, and so to understand the ‘big picture’ of how the Universe has evolved and is continuing to expand. The cosmic microwave background (CMB) is our only direct probe of conditions in the early Universe: it provides information about both the geometry and the contents of the Universe at the time the CMB was produced (i.e. $\sim 400\,000$ years after the big bang). By this time, nucleosynthesis had occurred and ordinary matter was primarily in the form of hydrogen and helium atoms. In this chapter we will explore the physics of how particle composition evolved in the *very* early Universe to reach this state.

Objectives

Working through this chapter will enable you to:

- describe the most important particle interactions that took place in the early Universe
- explain how the changing conditions in the early Universe determined the rates of key interaction processes
- summarise the timeline of the early Universe, from the big bang to primordial nucleosynthesis
- explain the importance of two key parameters – the baryon-to-photon ratio and the neutron-to-proton ratio – for the elemental abundances of the material from which the first stars and galaxies formed
- summarise the nuclear reaction processes that took place in the early Universe.

8.1 Particle interactions in the early Universe

An understanding of particle interactions in the early Universe is necessary to fill in the first steps of a timeline for how the present-day Universe came into being. Modelling these interactions also provides a further tool for testing cosmological models: the elemental abundances in the first stars and galaxies depend critically on conditions at earlier times, and astronomers can measure these abundances by searching for and studying the oldest regions that telescopes can access.

- Why weren’t hydrogen and helium atoms present in the Universe immediately after the big bang?

- The early Universe was very hot and dominated by radiation. Nuclei and atoms cannot form in those circumstances, and even protons and neutrons break down into their constituent quarks in this environment.

Examining the earliest times in the Universe forces us to confront the limits of our knowledge. Conditions in the early Universe must have reached extremes far beyond anything that can be simulated in a lab, and cosmologists have no direct ways to test theoretical ideas of how the Universe began, or whether something – or *anything* – existed prior to the big bang. In this chapter we will focus mainly on the stages for which the physics is well understood, involving particles such as protons (p), neutrons (n), electrons (e^-) and photons (γ). But we also need to set the scene for the advent of nucleosynthesis by considering some more exotic, and less well-understood, processes.

One of the main limitations of our knowledge relates to the existence and nature of dark matter. We saw in previous chapters that, initially, radiation (and then later, matter) drove the overall rate of expansion of the Universe, and also that dark matter played a key role in the oscillations that determine the CMB structure. So dark matter is clearly an important actor in the evolution of the early Universe.

However, in this chapter we want to explore how atoms formed; to do that we need to focus on particle interactions via short-range forces. These key processes do not involve the gravitational force by which dark matter interacts, which means that we can set aside dark matter for now, and focus on particles and processes that are better understood. We will return to the evidence for, and nature of, dark matter in the next chapter.

8.1.1 Basics of early-Universe interaction physics

The particle families of the Standard Model, and the four fundamental interactions by which they interact, were introduced in Chapter 1. Before we move on to consider the main particle interaction processes of the early Universe, you should review your understanding of these particles and forces with the following exercise.

Exercise 8.1

Using your existing knowledge (supplemented by brief further research if needed), complete Table 8.1 to summarise the key properties of the four fundamental interactions. Note that there is no universal definition of ‘relative strength’, and so you may obtain somewhat different numbers from different information sources.

Table 8.1 Basic properties of the four forces. (To be completed.)

Interaction	Particles affected	Carrier particle	Relative strength
Strong			
Weak			
Electromagnetic			
Gravity			

All of the main particle families and forces are important in the early Universe but, as mentioned above, we can consider gravity somewhat separately from the other three forces because of its relative weakness in the context of interactions between individual particles. Electromagnetic, weak and strong nuclear interactions are the key processes leading to the formation of atoms in the dense early Universe.

There are a variety of different particle interactions that could, in principle, occur in most physical environments, where numerous types of particle might exist. The crucial question is which process or processes have the *most* impact on the relative proportions of different particle types at different times? We therefore want to think about **reaction rates** for different processes, and which factors control those rates.

- Which global properties were changing in the early Universe and are likely to be relevant to the rates at which particles interact?
- As the Universe expanded, its temperature and density decreased. Temperature determines the energy and velocity distributions of particles, and both particle speed and density will affect how often and how strongly particles interact.

We can write a standard expression for the rate, Γ , at which a particle will undergo interactions of a particular type without specifying (yet) which particles and type of interaction are involved:

$$\Gamma = n\sigma v \quad (8.1)$$

where n is the number density of the ‘target’ particle with which an incoming particle might interact, v is the typical relative speed of the interacting particles and σ is the **interaction cross-section**: a measure of the probability that the interaction will occur. We will consider σ in more detail shortly.

Equation 8.1 describes the rate of reaction per incoming particle, but it is often useful to consider the rate at which reactions will occur within a particular volume. This will depend on the density of both types of

particle involved, so the rate of interaction between two particle species, A and B, per unit volume is

$$R_{AB} = n_A n_B \sigma v$$

where n_A and n_B are the number densities for each species of particle, respectively.

- What units must σ have if Γ has units of s^{-1} (i.e. interactions per second) and R_{AB} has units of $s^{-1} m^{-3}$ (i.e. interactions per second per cubic metre)?
- Rationalising the other units tells us that σ must have units of m^2 .

The fact that the interaction cross-section σ has dimensions of area can be understood by analogy to the simplest type of two-body interaction in classical physics: scattering between two solid bodies. Such interactions depend on the geometric area of the target object (a larger target cross-section provides more opportunity for the scattering of incident particles to occur). For electromagnetic, weak and strong nuclear interactions, the effective area within which an interaction is likely depends on the range of the interaction and other factors, such as the energy of the particles.

Earlier in this section we noted that temperature is one of the important controlling parameters for the rate at which interactions take place in a particular environment. A key feature of the early Universe is that all of the different particle families, including radiation and neutrinos, were interacting so frequently that they were in a state of thermal equilibrium. This situation means that the particle energy distributions obey standard expressions for fermions and bosons, and the associated temperatures of matter and radiation are equal. Therefore, while thermal equilibrium still holds, we can use a single temperature value to characterise the state of the Universe at any particular time.

This temperature, T , describing both the matter and the radiation at a particular time, is a very important parameter for how matter evolved, because temperature determines a quantity known as the **interaction energy**. This quantity is the typical energy available for particle interactions and is given by:

$$E_{\text{int}} \approx k_B T \tag{8.2}$$

Particle energies are often expressed in units of eV (electronvolts), where $1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$. Use the following exercise to get a sense of the sorts of numbers that could apply to E_{int} in the very early Universe.

Exercise 8.2

Estimate the interaction energy, in units of J and GeV, when the temperature is 10^{14} K .

The available interaction energy is important for **pair production**, the process by which quarks, leptons and more massive particles were first produced in the early Universe. Pair production involves converting energy into mass (via the famous equation $E = mc^2$). When the typical interaction energy drops below the rest-mass energy of the particles that the interaction produces, then pair production of that type becomes very improbable. So, for example, because they are much less massive, electron–positron pairs can be produced at lower energies than any of the quark pairs.

Another useful way to assess the impact of changing density and temperature is to consider typical timescales for particle interactions. Recalling the definition of the interaction rate Γ for a particular process (Equation 8.1), the typical timescale for a particle to interact is $t_{\text{int}} = 1/\Gamma$.

- How would you expect Γ to change with time for typical particle interaction processes in an expanding Universe?
- The density and the velocity of particles will decrease with time, so we would expect Γ to decrease with time.

We can then ask whether, on the timescale t_{int} for some particular process, the Universe expands significantly. The simplest way to define a suitable expansion timescale for comparison is via the Hubble parameter. In Chapter 5, the Hubble time at the present epoch was defined as $t_{\text{H}} = H_0^{-1}$. Similarly, for the early Universe we can define an expansion time $t_{\text{exp}} = H(t)^{-1}$. From the relationship $H(t) = \dot{a}/a$, it can be shown that in the time $1/H(t)$ the scale factor increases by a factor of e (i.e. by a little less than three times). If the Universe nearly triples in size on this interaction timescale, then the probability of an interaction taking place is very low. Therefore, setting $t_{\text{int}} \approx t_{\text{exp}}$ provides a rough estimate of when particular interactions become improbable.

- If a particular type of interaction stops at a time $t_{\text{int}} \approx t_{\text{exp}}$, how are the interaction rate and the Hubble parameter at that time related to one another?
- Using the definitions of t_{int} and t_{exp} above, $\Gamma \approx H(t)$ at the time interactions stop.

There are, therefore, many types of interaction that are *possible* in the early Universe but become very *improbable* at later times. For example, at the particle scale, at early times there is thermal equilibrium between quarks (and later, neutrons and protons), leptons, photons and neutrinos. This equilibrium is achieved by frequent interactions that distribute energy between the different particle types. Once the interaction rate Γ for relevant processes is $\sim H(t)$ then, first, neutrinos (and later, photons) **decouple** from or **freeze out** of the cosmic ‘soup’. They are no longer in equilibrium with the other particles so travel across the Universe unimpeded by interactions. These decoupling processes also have implications for how the remaining matter (mainly neutrons, protons and electrons) evolves.

8.1.2 Key interaction processes

In the previous section we discussed particle interactions in fairly general terms. We will now review the most important types of interaction in the early Universe, in roughly chronological order of how matter evolved. Section 8.2 will then set out a full timeline for the evolution of the early Universe, in which all of these processes play crucial roles.

Two important rules that govern particle interactions are the conservation of **baryon number**, which applies to all particle interactions, and the conservation of **lepton number**, which applies to all interactions involving leptons. The sums of both of these numbers must remain the same before and after any interaction between particles. Baryons and antibaryons have baryon numbers of 1 and -1 , respectively. Similarly, leptons and antileptons have lepton numbers of 1 and -1 , respectively. Quarks and antiquarks have baryon numbers of $1/3$ and $-1/3$, respectively.

Pair production and annihilation

In the time immediately following the big bang, the first (ordinary) matter in the Universe was produced by pair production. Although this mechanism stopped being important for the overall composition of the Universe within a few seconds, it remains an astrophysically important process in extreme environments in the present-day Universe (for example, in the vicinity of black holes).

The earliest instances of the process included the production of pairs of oppositely charged quarks from gluons and/or photons, and the production of leptons (e.g. electron–positron pairs) from photons. Various other types of subatomic particle can also be generated in the same way.

Panel (a) of Figure 8.1 depicts the creation of an electron–positron pair, an interaction in which high-energy photons are the ‘reactants’. Panel (b) of the figure shows electron–positron **annihilation**, which is the opposite process to pair production; here, a particle interacts with its antiparticle and both are destroyed, with energy released as a result.

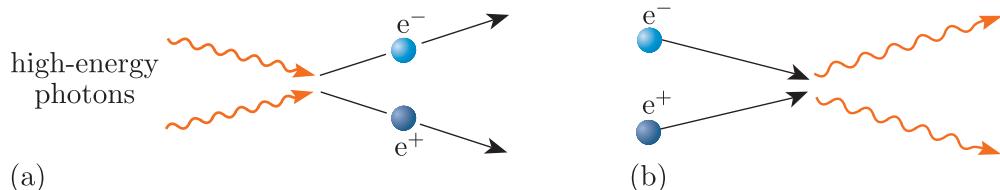


Figure 8.1 The opposing processes of (a) pair production from high-energy photons, and (b) annihilation.

Use the following exercise to consider the behaviour of different pair-production processes.

Exercise 8.3

Calculate the energies needed (in units of GeV) for the creation of a proton–antiproton pair and for an electron–positron pair. Which pair is easier to create?

During a period when the interaction energy is high compared to the masses of quarks, hadrons and leptons – as in the very early Universe – pair production and annihilation will both occur, but the former will dominate. When, as the Universe expands, the available energy drops then annihilation comes to dominate instead. This leads to one of the major conundrums of early-Universe physics: if particles are created in matter and antimatter pairs, why is the present-day Universe dominated by matter? In other words, why didn't all of the matter that was created in the early Universe subsequently annihilate? The answer to this question is not yet fully understood, but we will return to it in a later section.

Hadron production and dissociation

Quarks can bind together via the strong force to form hadrons, which include baryons and mesons. Baryons are hadrons containing three quarks, and include protons and neutrons, whereas mesons (of which the π meson is perhaps the most well known) contain a quark–antiquark pair.

Hadron formation involves the release of energy, much like the more familiar process of nuclear fusion. Conversely, the presence of a high density of energetic photons will dissociate hadrons into their constituent quarks. For example, the reversible process of proton formation – from two ‘up’ (u) quarks and one ‘down’ (d) quark – and dissociation is written as



The \rightleftharpoons symbol indicates that, in the early Universe, this process operated in both directions, with energies so high that when hadrons formed they rapidly dissociated again. Similar reactions apply to the formation and dissociation of neutrons, antiprotons and antineutrons. As the available energy decreased, the bound hadrons became stable and subsequently came to make up a major component of the cosmic soup.

Weak interactions

Unlike the strong force, weak interactions don't bind particles together, but they can change particles from one type to another. Many types of weak interaction occur in the very early Universe, but the most important for understanding the timeline of how matter evolved are those involving conversions between protons and neutrons. These reactions include processes known as **beta decay** (β -decay), and electron and positron capture. There are two forms of beta decay, termed beta-minus (β^-) and beta-plus (β^+) decay. You may have encountered these processes previously as important concepts in the physics of stars.

The term ‘beta decay’ comes from nuclear physics: in the context of radioactivity, emitted electrons or positrons are referred to as beta particles.

The conversion processes between neutrons and protons are summarised below:

Neutron–proton conversion reactions

β^- -decay involves the conversion of a neutron to a proton:



while β^+ -decay involves the conversion of a proton to a neutron:



where ν_e and $\bar{\nu}_e$ refer to (electron) neutrinos and antineutrinos, respectively.

Electron capture also involves the conversion of a proton to a neutron:



while **positron capture** involves the conversion of a neutron to a proton:



Because neutrons and protons are the building blocks of atoms, the way that these reactions proceed – and the eventual balance between the number of neutrons and protons in the Universe – is important in determining the composition of matter. At the earliest times in the Universe, the weak interactions listed above could also proceed in the reverse direction but, as density decreased, the probability of neutrinos interacting with other particles became very low.

Thomson and Compton scattering

A wide range of scattering processes – in which particles exchange energy but do not bind together or change species – took place in the early Universe. These processes contribute to ensuring thermal equilibrium between different species, including photons.

The most important such process for the cosmological timeline is Thomson scattering, which was introduced in Chapter 1 and also discussed in Chapter 6. In this form of scattering, photons are deflected by the electromagnetic field of charged particles, such as electrons. Thomson scattering occurs in the limit where the photon energy is much lower than the mass energy of the particle ($h\nu \ll mc^2$), and the particle energy and photon frequency remain unchanged by the interaction.

The rate at which particles will undergo Thomson scattering is described by Equation 8.1. The cross-section for Thomson scattering depends inversely on the square of the mass of the scattering particle. So in an ionised plasma, scattering by *electrons* is the dominant process.

Figure 8.2 compares Thomson scattering, in panel (a), with a situation where the photon energy is comparable to the energy of the particle with which it is interacting ($h\nu \sim mc^2$; panel (b)). In the latter case, the particle can gain energy from the interaction and the photon frequency will decrease. This phenomenon is known as **Compton scattering**.

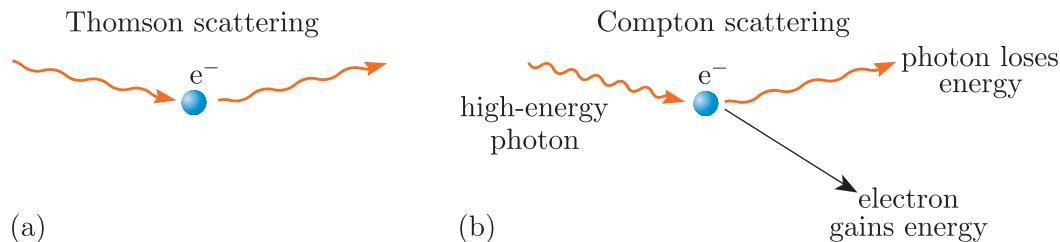


Figure 8.2 (a) Thomson scattering and (b) Compton scattering.

Finally, in situations where the particles have relativistic speeds, the reverse of the process shown in panel (b) can occur, causing photons to gain energy from the scattering process and increase their frequency. This is known as **inverse Compton scattering**, and is important in extreme environments in the present-day Universe, such as in the regions around black holes. However, it is not an important process for our discussion of the early Universe.

Nuclear fusion and photodisintegration

All of the elements in the present-day Universe formed via nuclear fusion. The majority of heavier elements were fused in stars, stellar mergers and explosions, but the fusion processes that operated in the very early Universe were the starting point for the composition of today's Universe.

The set of reactions by which fusion proceeded during the period known as **big bang nucleosynthesis** has some similarities to, but also some differences from, the fusion reaction chains that power the Sun and other stars. Both situations require high particle energies, enabled by high temperatures. But the presence of large numbers of free neutrons and photons in the early Universe, as well as the decrease in temperature as time elapsed, influenced how nucleosynthesis progressed.

The first important fusion reaction in the early Universe is the formation of deuterons, D^+ , which are nuclei containing one neutron and one proton:



Deuterons are a (heavier) form of hydrogen nuclei, and so are also sometimes written as ${}^2H^+$. You may previously have encountered deuterium, the atom consisting of a deuteron with a bound electron. This atom has an important use in nuclear reactors in the form of ‘heavy water’, D_2O , which is more dense than regular water and slows down the high-energy neutrons released by fission reactions, which serves to increase the likelihood of further fission reactions taking place.

Equation 8.8 shows the deuteron fusion reaction proceeding in both directions. Fusion reactions typically release energy, with the amount released corresponding to the binding energy of the reactants, as given by the difference in mass energy between the products and the reactants. However, as we saw with the dissociation of hadrons (e.g. Equation 8.3), if a *large* amount of energy is absorbed by a nucleus then it will break into its constituent parts once more. In other words, interactions with energetic photons can cause nuclei to disintegrate, a process known as **photodisintegration**. For a photon to cause a nucleus to disintegrate, its energy must be greater than the nuclear binding energy.

- How would you expect the balance between fusion reactions and photodisintegration to alter as the Universe expands?
- The density of radiation relative to matter decreases with time, so photodisintegration should become less important as the Universe expands.

Although nuclear fusion requires high energies, it could not proceed effectively at the very earliest times because of the high radiation density: nuclei that formed would rapidly break back down into their constituent hadrons. Fusion could only proceed once the photon density had dropped enough to reduce the typical photon energy. The following example quantifies this energy requirement in the context of a deuteron.

Example 8.1

Calculate the photon energy required to disintegrate a deuteron ($m_D = 3.3436 \times 10^{-27} \text{ kg}$) in units of J and MeV. Compare your findings to the interaction energy when the Universe had a temperature of $T = 10^{14} \text{ K}$: was deuteron photodisintegration feasible at this time?

Solution

In order for photodisintegration to take place, the incoming photon must have an energy greater than the binding energy of the atom. The binding energy of a nucleus is given by the **mass deficit**: the difference between the mass of the nucleus and that of the constituent particles.

The proton and neutron masses are $m_p = 1.6726 \times 10^{-27} \text{ kg}$ and $m_n = 1.6749 \times 10^{-27} \text{ kg}$. Therefore the mass deficit, Δm , is given by:

$$\Delta m = (m_p + m_n) - m_D = 3.9 \times 10^{-30} \text{ kg}$$

Using $E = \Delta m c^2$, we can calculate that this mass corresponds to a binding energy of $3.5 \times 10^{-13} \text{ J}$, which is $\sim 2.2 \text{ MeV}$.

We showed in Exercise 8.2 that at the time that the Universe's temperature was $T = 10^{14} \text{ K}$, the typical interaction energy was 8.6 GeV. Hence there was ample energy at that time for the photodisintegration of deuterons to occur.

Ionisation and recombination

The final pair of important processes to review are ionisation and recombination. Although ionisation can have multiple causes, it is photoionisation, which is caused by a particle absorbing a photon, that is of most importance in the early Universe. The ionisation of hydrogen can be written as



As earlier parts of this section have emphasised, the high radiation density and typical photon energy in the early Universe act to disrupt all types of interactions in which particles bind together. This is also the case for the binding of electrons to nuclei to form atoms. Hence when the first nuclei form, they exist as part of an ionised plasma containing oppositely charged nuclei and electrons; any atoms that form immediately dissociate again.

- How do local conditions determine the fraction of material that is ionised?
- For a predominantly hydrogen gas, the Saha equation (Equation 1.9 in Chapter 1) describes how the ratio of ionised protons and electrons to atoms depends on the temperature.

In the same way that photodisintegration decreases in importance as the Universe expands and photon energies decrease, so ionisation also becomes less favoured over time, and atoms can begin to form. In other words, the interaction described by Equation 8.9 comes to proceed mainly from right to left.

The binding of electrons to protons (and heavier nuclei) is known as recombination. This is a confusing term in the context of cosmology because the nuclei and electrons had not previously been combined. The term derives from experiments on plasmas, in which the nuclei and electrons were originally produced by ionising atoms, and it has become the accepted terminology in the study of the early Universe too, despite the potential for confusion.

8.2 A timeline towards the formation of atomic nuclei

The previous section reviewed the range of particle interactions of most importance for the evolution of matter and radiation in the early Universe. As time elapsed and the temperature of the Universe decreased, the dominant interactions changed. In this section we will set out a more detailed timeline, leading from the big bang to the point at which nuclear fusion became a significant process.

- Nuclear fusion reactions require extremely high temperatures. Why couldn't nuclear reactions take place immediately after the big bang?

- The high temperatures also meant there were very high densities of radiation. This first prevented quarks from forming protons and neutrons, and then prevented protons and neutrons from forming nuclei without immediately disintegrating.

It is therefore necessary to start at the beginning and follow the different stages by which particle composition evolved.

8.2.1 The very early Universe and inflation

The big bang itself, and conditions for the first fractions of a second afterwards, are very poorly understood. The **Planck time**, $t_P = 5.4 \times 10^{-44}$ s, is defined as the smallest time interval that can be measured, according to the theory of quantum mechanics. This means that we do not have any consistent physical theories that could be used to explain any processes occurring before t_P has elapsed.

Another reason that conventional theory fails at the very earliest times is because at the highest energies (well beyond the experimental conditions available for laboratory tests), the strong, weak and electromagnetic interactions are theorised to become a single interaction. Theories that explain this behaviour are known as **grand unified theories** (or GUTs), but a generally accepted GUT does not currently exist.

The idea of grand unification is motivated by the experimentally verified theory of **electroweak unification**, in which the weak and electromagnetic forces can be described as a single interaction at energies above ~ 250 GeV. The energy scale on which grand unification is predicted to occur is very much larger: $\sim 10^{16}$ GeV. When the age of the Universe was about 10^{-36} s, the high-energy conditions under which the strong and electroweak interactions are thought to have been unified came to an end.

Although the typical interaction energies at that time were far in excess of those in laboratory experiments, physicists do have a theoretical framework within which some predictions about this era of cosmic history can be made. It is predicted that for the brief period known as **inflation**, the scale factor of the Universe, $a(t)$, expanded exponentially (rather than more gradually, as measured observationally for much later times).

The theory of inflation was proposed by the physicist Alan Guth, and relates to the quantum field behaviour of empty space (the vacuum), which is proposed to have a non-zero associated energy. This idea of the energy density of empty space also underpins many theories of dark energy, to explain the acceleration of the Universe's expansion at the present time.

The inflationary era is thought to result from the physical state of empty space at the end of grand unification. These conditions allowed a much higher energy density of the vacuum than there is at the present day, leading to a dramatic, rapid expansion of space. Figure 8.3 provides an example of how the scale factor, a , *could* have evolved in the very early Universe.

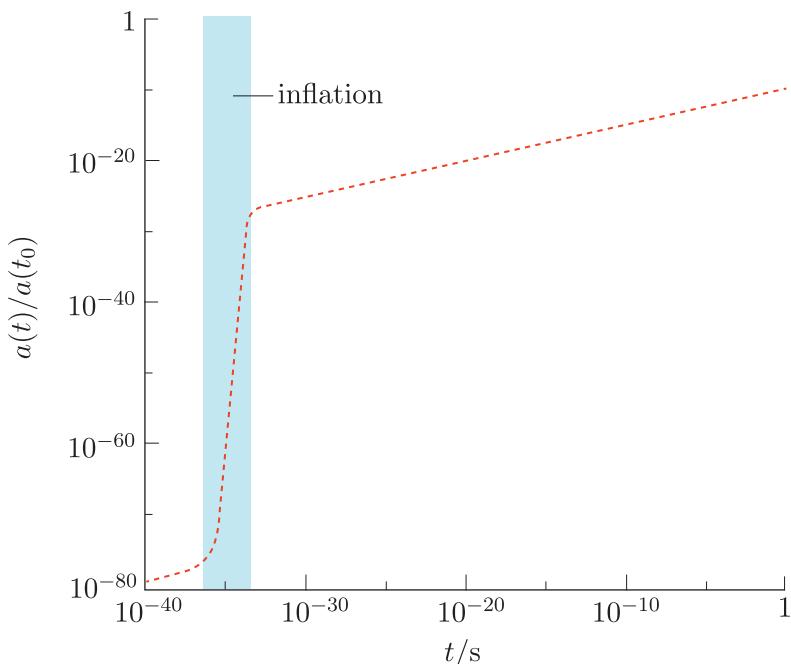


Figure 8.3 Possible evolution of the scale factor with time, including a period of inflation. Note that the numerical values shown here are not well determined.

The idea that space could expand by many tens of orders of magnitude in a fraction of a second does not immediately seem very plausible. But the reason some form of inflation is considered such a promising theory is that it resolves several hard-to-explain challenges in cosmology. These ideas will be discussed later in the module, but here we will move on to focus on the (much better-understood) physics of the next stages of the early Universe timeline.

As the period of inflation came to an end, the energy of the vacuum must have dropped to the level we see today. As the vacuum transitioned to its ‘true vacuum’ state (a more stable, lower energy state), energy was released and pair production took place. This process produced all types of quarks, antiquarks, leptons and antileptons. According to the inflationary model, the vast majority of particles in the Universe were created from the energy released as inflation came to an end.

When the temperature dropped sufficiently, the electromagnetic and weak interactions became separate; prior to this, there was no distinction in the mediating bosons. After this ‘electroweak transition’, weak and electromagnetic interactions proceeded as they do in present-day lab conditions, mediated respectively by W and Z bosons, and by photons. The constituents of the Universe immediately after this time continued to be all types of quarks and leptons and their antiparticles, as well as photons.

8.2.2 Neutrino decoupling and the baryon-to-photon ratio

When the temperature of the Universe dropped further, around 10^{-5} seconds after the big bang, quarks were able to become bound into hadrons without the process being disrupted. Although many types of hadron were formed in this way, only two were stable enough to have had a long-lasting effect on the composition of the Universe: the proton and the neutron. Both of these particles are baryons (i.e. they each contain three quarks).

The process by which the early Universe came to contain large numbers of baryons (and not antibaryons) is known as baryogenesis. How an imbalance was created – so that some fraction of the quarks that originally formed did *not* subsequently annihilate with antiquarks – is not fully understood, and must require physics beyond the Standard Model; some such theories associate baryogenesis with the GUT era, while others associate it with the separation of the weak and electromagnetic interactions.

At the start of what could be termed the hadron era, hadrons, leptons and photons were all in thermal equilibrium and undergoing frequent interactions and exchanges of energy. The interactions that were important for this thermal equilibrium were neutrino scattering processes involving electrons and positrons; for example, $\nu_e + \bar{\nu}_e \rightleftharpoons e^- + e^+$. As suggested by Equation 8.1, the rate of these scattering interactions, Γ_ν , is related to the number density of neutrinos, which decreases faster than the Hubble parameter during the radiation-dominated era.

When the interaction energy dropped to ~ 2.5 MeV, the neutrinos decoupled from the other particles. Essentially, the Universe became transparent to these primordial neutrinos, which could travel for billions of years without undergoing an interaction.*

Shortly after neutrino decoupling, when the age of the Universe was about 1 second, another important change took place: electrons and positrons have a mass energy of 0.511 MeV, and so when the interaction energy dropped to ~ 1 MeV the production of e^\pm pairs was no longer energetically favourable. As the temperature fell further, electrons and positrons began to disappear because e^\pm pairs were no longer being created, but the annihilation reaction ($e^+ + e^- \rightarrow \gamma + \gamma$) continued. As a result, the number of electrons and positrons decreased dramatically.

- How would the rapid annihilation of large numbers of electron–positron pairs affect the other constituents of the Universe at that time?

*This implies that there should be a cosmic neutrino background, analogous to the CMB. It would be another useful test for cosmological theory and early-Universe physics if we could observe such a background but, unfortunately, the very low interaction cross-section for neutrinos means that this is beyond the capability of current detectors.

- The e^\pm annihilation process produces large numbers of photons. The baryons, remaining electrons and radiation are in thermal equilibrium at this time, and so energy will be shared between them; the neutrinos will not gain any energy because they are no longer coupled to the other particles (including photons).

It is at this time – still only around 1 second after the big bang – that the value of a crucial cosmological parameter was set: the **baryon-to-photon ratio**, η . The nuclear reactions that form helium and heavier elements – and so the cosmic abundances that we can measure to test theories of early-Universe physics – all depend on this parameter.

The baryon-to-photon ratio

$$\eta = \frac{n_b}{n_\gamma} \quad (8.10)$$

where n_b and n_γ are the number densities of baryons and photons, respectively. Because the volumes over which the particles are distributed are the same, η is equivalently just the ratio of the total *numbers* of baryons to photons.

The photon density of the Universe is dominated by CMB photons – starlight and other photon-producing processes cannot compete with the number of photons produced in the early Universe. Hence n_γ can be measured from the present-day CMB. The baryon density is harder to measure, because matter in the present-day Universe is concentrated into galaxies and clusters of galaxies with large, quite empty voids between them. Nevertheless, n_b can be estimated reasonably well by observations on the scale of galaxy clusters and larger.

Example 8.2

Currently, the temperature of the CMB is 2.73 K and the baryon density is around 0.3 baryons per cubic metre. Determine the present-day number density of photons, and so estimate the value of η . (*Hint:* for a black-body distribution of photons, the mean photon energy is $E_\gamma \approx 3k_B T$.)

Solution

We will first determine the energy density of CMB radiation, ϵ .

The energy density of a relativistic gas is related to its pressure, $P = (1/3)\epsilon$, and via Equation 1.7 (in Chapter 1) to its temperature, $\epsilon = aT^4$, where a is the radiation constant. Hence at the present time:

$$\epsilon = (7.566 \times 10^{-16} \text{ J m}^{-3} \text{ K}^{-4}) \times (2.73 \text{ K})^4 = 4.20 \times 10^{-14} \text{ J m}^{-3}$$

The typical number of photons per cubic metre is given by $n_\gamma = \epsilon/E_\gamma$, where E_γ is the typical photon energy. Using the approximation

$E_\gamma \approx 3k_B T$, we therefore find that:

$$\begin{aligned} n_\gamma &= \epsilon/E_\gamma \\ &= (4.20 \times 10^{-14} \text{ J m}^{-3})/(3 \times 1.381 \times 10^{-23} \text{ JK}^{-1} \times 2.73 \text{ K}) \\ &\approx 3.7 \times 10^8 \text{ m}^{-3} \end{aligned}$$

In other words there are, on average, more than a third of a billion CMB photons per cubic metre. Given the present baryon density of $\sim 0.3 \text{ m}^{-3}$, the baryon-to-photon ratio, η , can be estimated at 8.1×10^{-10} , or $\sim 10^{-9}$.

As the example above demonstrates, it is estimated that at the present day there are approximately 1 billion photons for every baryon.

- Does η evolve with cosmic time? Briefly explain your answer.
- No. Baryon number is conserved in all particle interactions (see Section 8.1.2), and although photon production and annihilation does take place in stars and galaxies, the numbers are negligible compared with the total number of CMB photons produced in the early Universe.

The baryon-to-photon ratio must also be directly related to the baryon density parameter, Ω_b , which, as you have seen in previous chapters, can be constrained by observations of the CMB. The relationship to the present-day density parameter can be written as

$$\eta = \frac{\rho_{c,0}\Omega_{b,0}}{\langle m \rangle n_{\gamma,0}} \quad (8.11)$$

where $\rho_{c,0}$ is the present-day critical density and $\langle m \rangle$ is the mean particle mass. This relation is slightly sensitive to other cosmological parameters, because the mean baryon mass in the present-day Universe – although very close to the proton mass – is affected by the nucleosynthesis processes that occur a little later in the history of the Universe.

8.2.3 The neutron-to-proton ratio

Both the decoupling of neutrinos and electron–positron annihilation had a substantial effect on the range of weak interactions that could take place in the first seconds after the big bang. When all particles were abundant and in thermal equilibrium, neutrons could convert to protons and vice versa by interacting with neutrinos, antineutrinos, electrons and positrons, according to the following equilibrium reactions:



The ratio of neutrons to protons, n_n/n_p , turns out to be a crucial parameter for cosmic nucleosynthesis. Its value at early times is set by statistical equilibrium between the interactions of Equations 8.12 and 8.13. The probability p that particles at equilibrium will occupy a given state within these reactions depends on the energy of the state and the

gas conditions via the **Boltzmann factor**:

$$p_x \propto \exp\left(-\frac{E_x}{k_B T}\right) \quad (8.14)$$

where x represents a particular particle state. For example, if p_x is the probability that the particle is in the state of being a neutron, then E_x is the rest-mass energy of a neutron, $m_n c^2$.

The ratio of neutrons to protons in a situation of statistical equilibrium is therefore given by:

$$\frac{n_n}{n_p} = \exp\left[-\frac{(m_n - m_p)c^2}{k_B T}\right] \quad (8.15)$$

This expression may look familiar: the Saha equation describing ionisation fractions introduced in Chapter 1 is derived in a similar way. You can now test your understanding of it by completing Exercise 8.4.

Exercise 8.4

Show that the neutron-to-proton ratio is $\sim 1/5$ immediately after electron–positron annihilation (at $t \sim 1$ s), when the interaction energy $E \approx 0.8$ MeV.

When far fewer electrons and neutrinos were available for frequent interactions, the neutron-to-proton ratio no longer evolved with temperature according to Equation 8.15. Instead, the only process capable of changing baryon type was the β^- -decay of free neutrons (Equation 8.4). Therefore, the neutron-to-proton ratio slowly decreased as neutrons were converted to protons. The neutron decay lifetime, $\tau_n = 887$ s, is therefore an important cosmological parameter, and the total number N of neutrons in the Universe evolved according to:

$$N(t) = N(t_{\text{init}}) \exp\left(-\frac{t}{\tau_n}\right) \quad (8.16)$$

where t_{init} is taken as a starting time at which N is known.

Exercise 8.5

If free-neutron decay continued indefinitely, calculate the age of the Universe for which the neutron-to-proton ratio would reach a value of $n_n/n_p \approx 0.05$ (i.e. 1 neutron for every 20 protons), assuming $n_n/n_p = 1/5$ at $t = 1$.

(Hint: you may make the simplifying assumption that the number of protons remains constant, because there are already very many more neutrons than protons present when the decay starts.)

8.3 Big bang nucleosynthesis

We have now reached the point in the history of the Universe where nuclear fusion processes become important. This stage is known as big bang nucleosynthesis (or **primordial nucleosynthesis**), and it is thought to have taken place over a timescale of around 20 minutes.

The physics of big bang nucleosynthesis explains why the Universe is composed primarily of hydrogen and helium, and also provides astronomers with some direct tests of conditions of early times. The key parameters we considered in the previous section, η and n_n/n_p , strongly influence what happened next.

8.3.1 Deuterium formation

The first nuclear fusion reaction to occur in the Universe was briefly introduced in Section 8.1.2: deuterium nuclei (deuterons) formed via the fusion of a neutron and a proton, which also resulted in the release of energy (Equation 8.8). In that section we saw that the competing process of photodisintegration also plays an important role at early times: deuterons can only begin to form in large numbers when the temperature drops sufficiently to suppress photodisintegration.

Example 8.1 showed that photons of energy $\sim 2.2 \text{ MeV}$ are needed to disintegrate deuterons, which might suggest that by the time of electron–positron annihilation ($k_B T \approx 0.8 \text{ MeV}$), fusion could proceed effectively. However, we have also seen that the density of photons relative to baryons is of the order 10^9 ; this means that we need to consider the *distribution* of photon energies more carefully. Even if only a small fraction of the photons have sufficient energy to cause photodisintegration, this quantity might be enough to disrupt a large proportion of fusion interactions.

The distribution of photons according to frequency follows a black-body spectrum (see Section 1.3.2 in Chapter 1, and Figure 6.4). The number density of black-body photons with frequencies between ν and $\nu + d\nu$ is given by:

$$n(\nu, T) d\nu = \frac{8\pi}{c^3} \frac{\nu^2}{\exp(h\nu/k_B T) - 1} d\nu \quad (8.17)$$

Calculating the number of photons in the high-energy tail of the distribution requires integration of this function, which is not algebraically trivial; the following example uses results from numerical integration.

Example 8.3

Figure 8.4 plots the fraction of photons with energy greater than 2.2 MeV in a black-body distribution (i.e. $n(E > 2.2 \text{ MeV})/n_{\text{tot}}$) as a function of temperature.

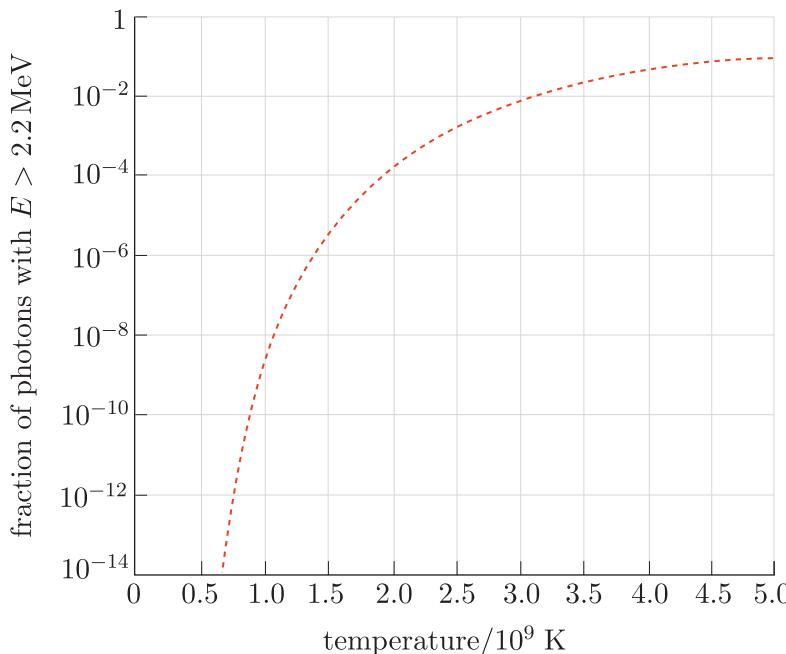


Figure 8.4 Fraction of black-body photons with $E > 2.2 \text{ MeV}$ as a function of temperature.

- (a) Based on this graph, could nuclear fusion proceed effectively immediately after electron–positron annihilation (when the interaction energy $E_{\text{int}} \approx 0.9 \text{ MeV}$)?
- (b) If not, then to what level must E_{int} drop before deuterons could survive?

Solution

- (a) In order for deuterons to survive, the number of photons energetic enough to disintegrate them must be fewer than the number of baryons. In other words,

$$n_{\gamma > 2.2} < n_b$$

where $n_{\gamma > 2.2}$ is the number density of photons with energy above 2.2 MeV .

If we divide both sides by the total number density of photons, n_{tot} , and take $f = n(E > 2.2 \text{ MeV})/n_{\text{tot}}$ (as plotted in Figure 8.4) to be the fraction of photons with energy above 2.2 MeV , then the condition for deuteron survival becomes:

$$f < \eta$$

In other words, the fraction of sufficiently energetic photons must be lower than the baryon-to-photon ratio.

An interaction energy of $E_{\text{int}} = 0.9 \text{ MeV}$ corresponds to a temperature of $T = E/k_B = 10^{10} \text{ K}$. This temperature is well beyond the right-hand edge of the graph, so it is clear that the value of f at this temperature will be higher than that for the highest plotted

temperature of 5×10^9 K, where $f \rightarrow 0.1$. In Example 8.2 you saw that $\eta \approx 10^{-9}$ so the condition for deuteron survival is not met, i.e. deuteron fusion cannot proceed effectively immediately after electron–positron annihilation.

- (b) Deuteron fusion can only proceed when $f < \eta$. The graph shows that f is lower than $\eta \approx 10^{-9}$ at temperatures $T < 1 \times 10^9$ K. This corresponds to an interaction energy $E_{\text{int}} \approx 0.09$ MeV.

The previous example has demonstrated why η is such an important parameter for how the early Universe evolved. If the value of η was substantially different then the start of nucleosynthesis would have occurred at a different temperature, and so significantly earlier or later. You may wish to explore further the distributions of photon energy at different temperatures via the following exercise.

Exercise 8.6

This is an *optional* Python exercise.

By adapting examples of numerical integration from the Python toolkit or earlier week-long practical activities, calculate the fraction of photons with $E > 2.2$ MeV for temperatures of $T = 10^9$ K and $T = 5 \times 10^9$ K, and check that your results are in agreement with Figure 8.4.

(*Hint:* to estimate n_{tot} you can assume that a reasonable upper limit for photon energy is 20 MeV.)

The overall impact of photodisintegration was to delay the onset of fusion and the formation of deuterons. The first nuclei formed at a lower temperature than would have been the case if the photon density was lower. This had an important knock-on effect for the subsequent evolution of matter. Further nuclear reactions – to form heavier elements such as helium – could only proceed once a supply of deuterons was available, so this stage of nucleosynthesis is referred to as the **deuterium bottleneck**.

- How would the subsequent evolution of matter in the Universe have been affected if deuteron formation happened at a higher or lower temperature?
- All nuclear reactions are strongly temperature-dependent, so further reactions – and the length of time over which they could operate – would differ significantly had the bottleneck occurred under different conditions. This would have the effect of altering relative abundances of elements.
- Which process in the early Universe could have resulted in a substantially different value of η , and hence a different timeline for nucleosynthesis?

- The poorly understood process of baryogenesis determined the number density of baryons in the early Universe, so strongly influenced the timeline of nucleosynthesis.

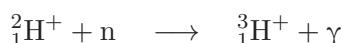
8.3.2 Helium formation

There is copious observational evidence that the Universe contains a large quantity of helium (He) – indeed the name of the element comes from its initial discovery in the outer regions of the Sun (in Greek mythology, Helios is the god of the Sun). Although helium is produced in stars, by far the bulk of the He in the present-day Universe is understood to have been synthesised in the first few hundred seconds of the early Universe.

Reaction pathways towards helium

There were multiple routes to helium production in the early Universe, with two different isotopes initially being produced: $^3_2\text{He}^{2+}$ and $^4_2\text{He}^{2+}$. The dominant processes are set out below. Note that, as indicated by the charges on each species, it is important to recall that here we are discussing the fusion of *nuclei* rather than atoms, because it is still too hot at this time for electrons to bind to the nuclei.

We can start by considering the formation from deuterium ($^2_1\text{H}^+$) of an even heavier isotope of hydrogen known as tritium ($^3_1\text{H}^+$) via one of the following reactions:



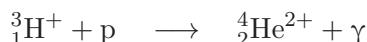
or



This newly formed tritium could then undergo further reactions to form helium-4:



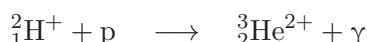
and



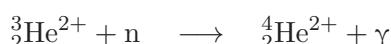
Helium-3 could also be formed from the same ingredients as formed tritium:



and



These helium-3 nuclei can then further react to form helium-4:



and



If you have previously studied nuclear fusion in stars then you may have noticed that the reactions to form helium in the early Universe differ from the proton–proton chain operating in the core region of the Sun and other stars. The reason is the high abundance of free neutrons in the early Universe (unlike in the centre of stars), which provide an additional – and in some cases easier – route to the assembly of heavier isotopes and nuclei.

- Why is it easier for a proton or a deuterium nucleus to fuse with a neutron, rather than with another identical proton or deuterium nucleus?
- Protons and deuterium nuclei are both positively charged, and so fusion reactions require the electromagnetic (Coulomb) repulsion between them to be overcome; it requires less energy for a proton or deuteron to get sufficiently close to a free neutron (if present) to fuse than for it to get close to another proton or deuteron.

Helium abundance and n_n/n_p

We have seen that the presence of free neutrons is crucial for the reaction pathways involved in fusion into helium. As a result of the reactions described in this section, almost all of the free neutrons available at early times became locked into helium. This means that the cosmic abundance of helium is closely linked to the neutron-to-proton ratio at the time of nucleosynthesis.

Let's consider the **mass fraction** of helium, Y , for a representative sample of the Universe (i.e. a region large enough to have a typical composition). Y is defined as the fraction of mass contained as helium relative to the total mass of baryons. We will see shortly that the contribution of elements heavier than helium to the total mass is very small, and so the total baryonic mass can be approximated as the sum of all of the hydrogen and helium within the region. Finally, we can make the simplifying assumption that all of the helium is helium-4.

The mass fraction of helium can then be estimated as

$$Y = \frac{N_{\text{He}}m_{\text{He}}}{N_{\text{H}}m_{\text{H}} + N_{\text{He}}m_{\text{He}}} \quad (8.18)$$

where N_{He} and N_{H} are the numbers of helium and hydrogen nuclei, and m_{He} and m_{H} are the masses of helium and hydrogen nuclei.

The next example examines the influence of the neutron-to-proton ratio on helium abundance.

Example 8.4

By expressing Equation 8.18 in terms of the numbers of protons and neutrons, show that the helium mass fraction in the Universe, Y , depends only on the neutron-to-proton ratio, n_n/n_p , and calculate the value of Y assuming that n_n/n_p is 0.16 at the start of nucleosynthesis.

Solution

We can start by making the simplifying assumption that $m_{\text{He}} = 4m_{\text{H}}$, which means Equation 8.18 can be written as

$$Y = \frac{N_{\text{He}} 4m_{\text{H}}}{N_{\text{H}} m_{\text{H}} + N_{\text{He}} 4m_{\text{H}}} = \frac{4N_{\text{He}}}{N_{\text{H}} + 4N_{\text{He}}} \quad (8.19)$$

We can now relate the numbers of helium and hydrogen nuclei to the numbers of protons and neutrons. Here we need to make another simplifying assumption, which is that all of the deuterium and tritium produced reacted to form helium-4. This allows us to assume that all of the neutrons are in helium, so that the number of helium nuclei will be half the number of neutrons because there are two neutrons in each helium nucleus (i.e. $N_{\text{He}} = N_{\text{n}}/2$).

Then the number of hydrogen nuclei will be the total number of protons minus the number of protons locked up in the helium nuclei, so

$$N_{\text{H}} = N_{\text{p}} - 2N_{\text{He}} = N_{\text{p}} - N_{\text{n}}.$$

Substituting in the expressions we've derived for N_{H} and N_{He} , we obtain:

$$Y = \frac{2N_{\text{n}}}{N_{\text{n}} + N_{\text{p}}}$$

To find an expression in terms of the neutron-to-proton ratio we can divide both the numerator and denominator by N_{n} to get:

$$Y = 2 \left(\frac{1}{1 + N_{\text{p}}/N_{\text{n}}} \right) = 2 \left(\frac{1}{1 + n_{\text{p}}/n_{\text{n}}} \right)$$

where the final step involved replacing the ratio of the total number of neutrons and protons with the (equivalent) ratio of their number densities.

Finally, using the given value of $n_{\text{n}}/n_{\text{p}} = 0.16$, we find that $Y = 0.28$, so our calculation predicts that helium should make up 28% of the baryonic mass of the Universe.

It is a remarkable success of the big bang model that the observed helium abundance in interstellar gas ($\sim 25\%$) is in good agreement with the predictions of early-Universe physics and big bang nucleosynthesis. In the next chapter we will compare predictions and observations of cosmic abundances more carefully.

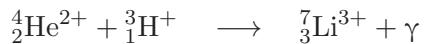
8.3.3 Formation of heavier nuclei

The major product of primordial nucleosynthesis was helium-4. The fact that nucleosynthesis did not progress further and produce large quantities of nuclides with higher mass numbers is due to two factors.

First, the rate at which two nuclei will fuse together depends very strongly on the temperature, and higher temperatures are required to fuse nuclei of higher atomic number. Because deuterons are easily photodisintegrated, the process of nucleosynthesis could only start once the deuterium bottleneck had been overcome, by which time the temperature of the

Universe was relatively low. As a consequence, the rate of fusion reactions that involved nuclides other than hydrogen and helium would have been very low.

The second factor is the lack of any stable nuclides with mass numbers of 5 or 8, which meant that helium-4 could not react either with itself or with the two most abundant species – protons and neutrons. This hurdle could, however, be overcome by reactions involving tritium or helium-3:



although because it is highly unstable, ${}^7_4\text{Be}^{4+}$ decays quickly to lithium-7:



Lithium nuclei can also undergo fusion with protons, but the result is the destruction of the lithium nucleus and the formation of two nuclei of helium-4:



The quantity of lithium produced from the above reactions (the **yield**) was small – the proportion of baryonic matter made up of lithium, or the lithium mass fraction, is $\sim 10^{-9}$, justifying the decision immediately prior to Example 8.4 to neglect heavier elements in our estimation of helium abundance! Nevertheless, predictions for the abundance of lithium are astrophysically interesting, and we will discuss in the next chapter how they have been used to test conditions in the early Universe.

8.4 Summary of Chapter 8

- The big bang model predicts that the early Universe was extremely hot and dense. At the very earliest times, conditions were sufficiently extreme that current physical theories (e.g. the Standard Model and quantum field theory) cannot fully describe them. A (poorly understood) brief period of **inflation** is thought to have taken place during which the Universe expanded rapidly.
- A ‘soup’ of many different particle types was present after inflation; the particles (e.g. quarks, antiquarks, leptons, antileptons, photons, gluons, and W and Z bosons) initially interacted very frequently with each other.
- The temperature and density of the Universe (for both matter and radiation) decreased with time, with the types of particle interaction taking place determined by the **interaction energy**, $E_{\text{int}} \approx k_B T$ (Equation 8.2), as well as the relative quantities of different types of particle.

- Important processes included:
 - **pair production** (of quarks and leptons) and **annihilation**
 - hadron production (binding of quarks via the strong force)
 - **beta decay**
 - **electron capture and positron capture**
 - Thomson scattering
 - a variety of nuclear fusion reactions
 - the **photodisintegration** of nuclei.
- The sequence of important particle interactions is summarised in Figure 8.5 (overleaf).
- The **reaction rate** for a particle of a particular type is given by $\Gamma = n\sigma v$ (Equation 8.1), where n is the number density of the target particle, σ is the **interaction cross-section**, and v is the relative speed of the interacting particles.
- Interaction rates typically decline as the Universe expands, and so the time at which a particular interaction becomes improbable can be estimated by equating Γ and the Hubble parameter, $H(t)$.
- Thermal equilibrium between different particle species was first broken when neutrinos **decoupled** from the other particles in the cosmic soup, which occurred after hadron production and was quickly followed by rapid annihilation of electrons and positrons.
- Neutrino decoupling and electron–positron annihilation set the value of the **baryon-to-photon ratio**, η , which is related to the present-day baryon density parameter via:

$$\eta = \frac{\rho_{c,0}\Omega_{b,0}}{\langle m \rangle n_{\gamma,0}} \quad (\text{Eqn 8.11})$$

- The baryon-to-photon and neutron-to-proton ratios had an important influence on the process of **big bang** or **primordial nucleosynthesis**, controlling how deuterium production evolved.
- Once a significant quantity of deuterons had been fused (the **deuterium bottleneck**), helium production rapidly consumed all of the remaining free neutrons. A small amount of lithium and beryllium were produced before the Universe became too cool to provide sufficient interaction energy for further nuclear reactions.

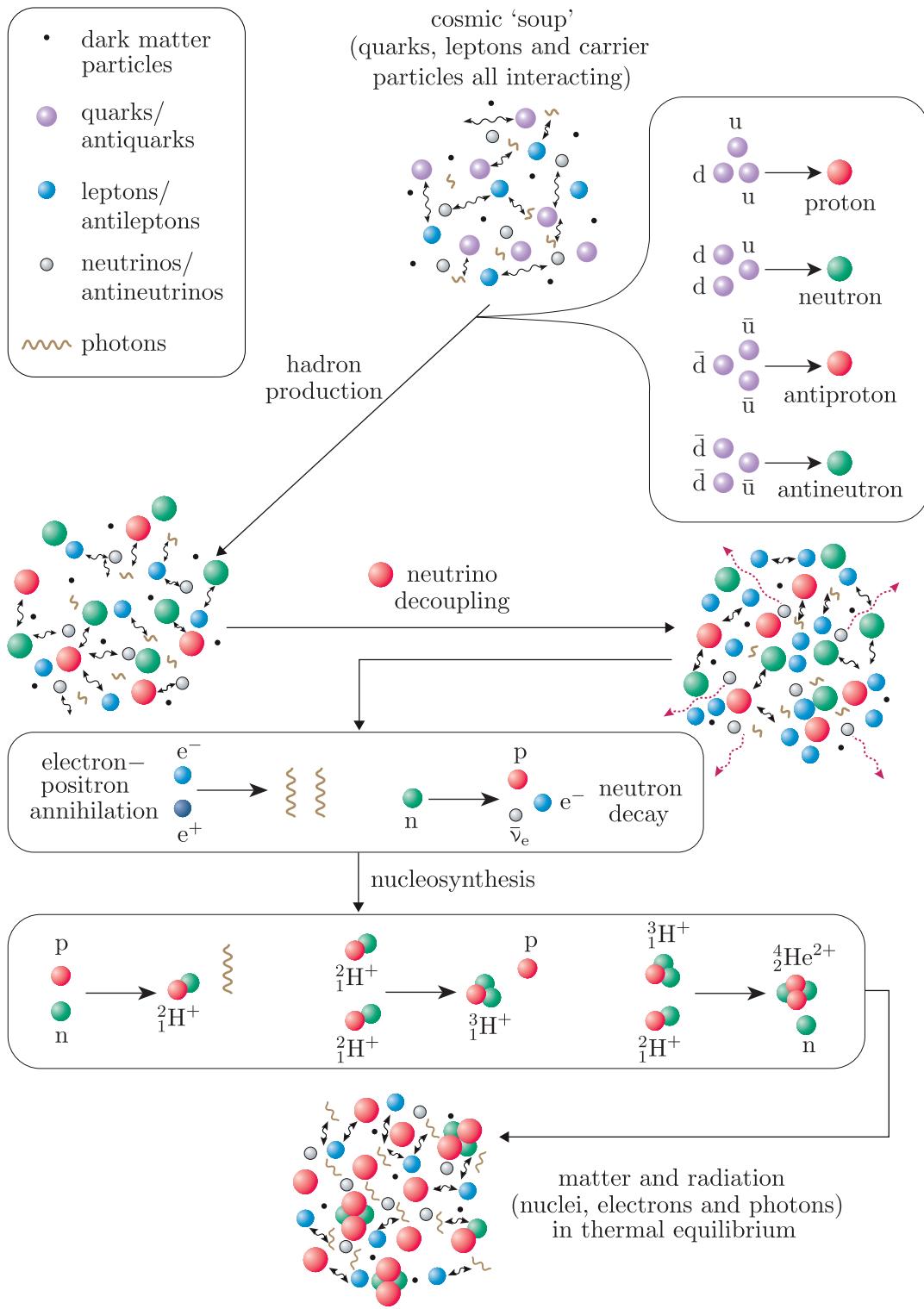


Figure 8.5 The evolution of the main particle species over the first 1000 seconds of the Universe's timeline. Note that this representation is not intended to show accurate relative quantities for different species (and does not include all the types of particles present immediately after the big bang).

Chapter 9 Early Universe physics meets observations

The previous chapter described a timeline extending from the creation of the very first particles to the production of the first elements via nucleosynthesis. This timeline is controlled by the expansion of spacetime and the resulting decrease in temperature and particle density with time.

As the Universe continued to cool and expand, it became possible for nuclei to combine with free electrons to form the first atoms. This is a crucial point in the history of the Universe, because it is around this time that the cosmic microwave background (CMB) was produced, enabling us to observe directly the conditions of the early Universe. In this chapter we will consider how a range of observations provide tests of the theory of how the early Universe evolved, and of the relative contributions of different types of matter, including dark matter.

Objectives

Working through this chapter will enable you to:

- describe, and mathematically investigate, the interactions of matter and radiation at the time when the CMB was produced
- discuss how the observed abundances of chemical elements in the Milky Way and other galaxies provide evidence for the big bang model
- explain how the cosmic abundances of helium, deuterium and lithium can be measured, and the challenges and limitations in making these measurements
- explain the importance of baryon density, Ω_b , as a cosmological parameter
- describe and critically discuss the key evidence for a dominant quantity of non-baryonic (i.e. dark) matter.

9.1 From nuclei to atoms

9.1.1 Recombination

After the formation of baryons, the Universe remained ionised for around 250 000 years. While the temperature remained high enough for photoionisation to be a dominant process, nuclei and electrons could not remain stably combined.

Eventually, as the temperature dropped, the ionised plasma was able to transform into a neutral gas of atoms – this transition is known as recombination. As noted in Chapter 8, in the context of the early Universe this term is misleading, because it is the *first* time that atoms form.

- By analogy with the process of deuterium synthesis, what quantity other than the mean temperature will influence when recombination takes place?
- The baryon-to-photon ratio, η . If there are large quantities of photons in a particular environment then the high-energy tail of the photon distribution can still photoionise atoms. The importance of these rarer high-energy photons for the recombination process depends on the overall number of photons relative to baryons.

The balance between ionised and neutral gas is described by the Saha equation:

$$\frac{n_{\text{H}}}{n_{\text{p}} n_{\text{e}}} = \left(\frac{m_{\text{e}} k_{\text{B}} T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_{\text{B}} T}\right) \quad (9.1)$$

where Q is the ionisation energy for hydrogen.

- What simplifying assumption does Equation 9.1 make about the composition of the gas?
- That the gas is purely composed of hydrogen.

While we know from the previous chapter that deuterium, helium and lithium were also be present at this point in the history of the Universe, it is sufficient to make the simplifying assumption of a hydrogen-only gas in order to get a first estimate of recombination conditions.

- In a hypothetical universe with a significantly lower value of η , would recombination occur at an earlier or later point in time?
- If there were fewer baryons per photon, then the high-energy photons in the tail of the CMB distribution would have a proportionally greater ionising effect at the same temperature. Recombination would therefore be delayed until the mean temperature was lower.

The following example investigates this dependence in more detail.

Example 9.1

Rearrange the Saha equation to show that the ionisation fraction of the Universe

$$X = \frac{n_{\text{p}}}{n_{\text{p}} + n_{\text{H}}}$$

depends on η as well as temperature.

Solution

We start by using the result of Example 1.2 from Chapter 1:

$$\frac{1 - X}{X} = n_{\text{p}} \left(\frac{m_{\text{e}} k_{\text{B}} T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_{\text{B}} T}\right) \quad (9.2)$$

Other than temperature, the unknown quantity in this equation is the proton number density, n_{p} . We need to eliminate n_{p} by relating it to the quantity of interest, η .

Since $\eta = n_b/n_\gamma$ and $X = n_p/n_b$, it follows that:

$$\eta = \frac{1}{X} \frac{n_p}{n_\gamma} \quad (9.3)$$

Rearranging for n_p gives:

$$n_p = \eta X n_\gamma \quad (9.4)$$

and substituting this into Equation 9.2 gives:

$$\frac{1-X}{X} = \eta X n_\gamma \left(\frac{m_e k_B T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_B T}\right)$$

and so:

$$\frac{1-X}{X^2} = \eta n_\gamma \left(\frac{m_e k_B T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_B T}\right)$$

We now have an expression that relates ionisation fraction to η , T and n_γ .

But, as you saw in Example 8.2 in Chapter 8, the photon number density depends only on T – it is the photon energy density $\epsilon = aT^4$ divided by the mean photon energy ($\approx 3k_B T$) – and so the ionisation fraction expression can be written as:

$$\frac{1-X}{X^2} = \eta \left(\frac{aT^3}{3k_B} \right) \left(\frac{m_e k_B T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{k_B T}\right) \quad (9.5)$$

We have therefore shown that the ionisation fraction depends only on η and T (albeit in an algebraically complicated way!)

Note that in this context the symbol a denotes the radiation constant rather than the scale factor.

Recombination does not take place instantly, but a simple estimate for the time of recombination is to take the point at which half of the baryons are contained within atoms (i.e. $X = 0.5$).

Exercise 9.1

Show that if the baryon-to-photon ratio $\eta = 10^{-9}$, then recombination can take place when the temperature is around 3795 K. (*Hint:* you do not need to solve for T to demonstrate this.)

It is possible to estimate the redshift of recombination by assuming a relation between T and the scale factor, a . T is proportional to a^{-1} and so $T \propto 1+z$.

- Given the recombination temperature of $T_{\text{rec}} \approx 3795$ K, estimate the redshift at which recombination occurred. (You can assume that the present-day CMB temperature is $T_0 \approx 2.7$ K.)
- Using the same approach as Exercise 1.4 in Chapter 1 we can substitute values into

$$\frac{T_{\text{rec}}}{T_0} = \frac{1+z_{\text{rec}}}{1+z_0}$$

to find that $z_{\text{rec}} \approx 1400$.

More sophisticated approaches to this calculation account for the additional ionisation caused by photons released in the recombination process, and for recombination to excited states of hydrogen. These lead to slightly lower estimates of z_{rec} .

9.1.2 Decoupling of matter and radiation

We have now reached the point in the timeline of the early Universe where theory starts to connect directly to observations. Soon after recombination comes the point at which photons decouple from the baryonic gas and are able to travel unimpeded to our telescopes, i.e. the time of last scattering, when the Universe becomes transparent to radiation.

- In Chapter 6 you read about the concept of last scattering: the point at which a typical photon last undergoes Thomson scattering (i.e. interacts with a free electron). How would you expect the process of recombination to influence the scattering of photons?
- Recombination binds free electrons to atoms, and so greatly reduces the number of free electrons available to scatter photons.

Recombination therefore dramatically reduced the rate of Thomson scattering, so decoupling of matter and radiation occurred fairly soon after this point in time. Figure 9.1 shows these final two stages of early Universe particle evolution.

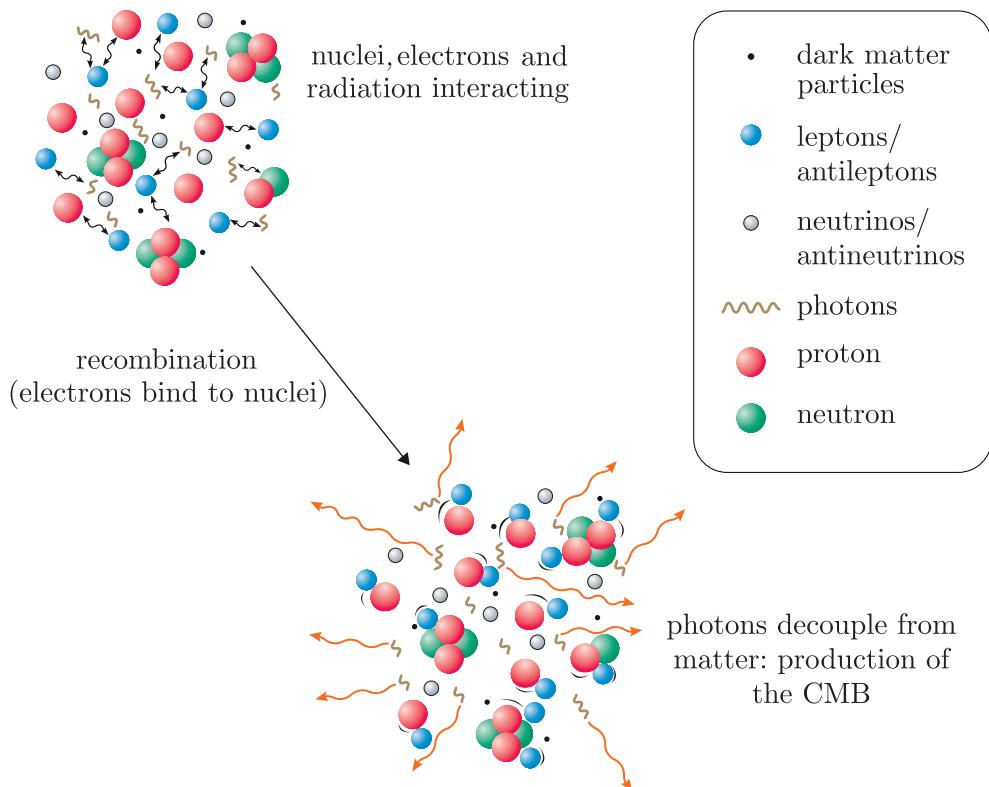


Figure 9.1 The evolution of particles and their interactions around the time of CMB production.

Decoupling is defined to be the point when the rate of photon–electron interactions drops below the rate at which the Universe expands, and is effectively the same point in cosmic history as the epoch of last scattering (z_{ls}). The situation is similar to the neutrino decoupling we considered in Chapter 8, and we can use the same approach of comparing the interaction rate, Γ , with the Hubble parameter, $H(t)$, explained in Section 8.1.1.

The interaction rate for Thomson scattering is given by:

$$\Gamma_e = n_e \sigma_T c \quad (9.6)$$

where σ_T is the Thomson cross-section.

The electron number density is highly dependent on the ionisation fraction, and so it changes with redshift. Γ_e can be related to the present-day baryon number density $n_{b,0}$ via:

$$\Gamma_e = X(z)(1+z)^3 n_{b,0} \sigma_T c \quad (9.7)$$

The Hubble parameter can also be expressed as a function of redshift instead of time, as follows:

$$H(z) = H_0 \sqrt{\Omega_{m,0}(1+z)^3} \quad (9.8)$$

and so equating $\Gamma_e(z)$ and $H(z)$ and rearranging leads to the following expression for the redshift of decoupling (or last scattering):

$$z_{ls} = \left[\frac{H_0 \sqrt{\Omega_{m,0}}}{X(z) n_{b,0} \sigma_T c} \right]^{2/3} - 1 \quad (9.9)$$

Equation 9.9 shows that it is necessary to know how the ionisation fraction depends on z to estimate the redshift at which decoupling takes place. (This makes sense, because interactions of photons with electrons are the most important interaction for how the radiation propagates.) The example below uses tabulated values of $X(z)$ to estimate z_{ls} .

Example 9.2

Table 9.1 lists the ionisation fraction $X(z)$ for different redshift values following the time of recombination, obtained from a modern recombination modelling code that accounts for the additional corrections mentioned at the end of Section 9.1.1.

Table 9.1 Tabulated values of hydrogen ionisation fraction $X(z)$.

z	$X(z)$
900	6.3×10^{-4}
1000	2.5×10^{-3}
1100	6.8×10^{-3}
1200	3.5×10^{-2}
1300	7.9×10^{-2}
1400	2.2×10^{-1}

Use this table, and Equation 9.9, to estimate the redshift of decoupling. Assume the present-day baryon density is 0.249 m^{-3} .

Solution

All of the parameters in Equation 9.9 apart from $X(z)$ are known, and redshift-independent. Since we only have tabulated values and not an expression to describe $X(z)$ we can't solve algebraically for z .

Instead we must substitute values of X into the equation, and see if there is a value that gives a self-consistent result. In other words, for each row in the table we can evaluate the right-hand side of Equation 9.9 and compare the resulting redshift to the corresponding value in that row of the table.

Using the *Planck* 2018 parameter values of $H_0 = 67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}$ (which can be expressed as $2.19 \times 10^{-18} \text{ s}^{-1}$ in SI units) and

$\Omega_{\text{m},0} = 0.3097$, and the provided value of $n_{\text{b},0}$, we can now evaluate z_{ls} for each row of the table and compare it to the input value.

It may be most efficient to write a short Python script to calculate the results, but this could also be done by hand. The results are shown in Table 9.2, rounded to two significant figures.

Table 9.2 z_{ls} calculation results.

z	$X(z)$	$z_{\text{dec}} (\text{output})$
900	6.3×10^{-4}	5300
1000	2.5×10^{-3}	2100
1100	6.8×10^{-3}	1100
1200	3.5×10^{-2}	370
1300	7.9×10^{-2}	210
1400	2.2×10^{-1}	110

It is clear from Table 9.2 that the calculation does not give self-consistent output at low and high redshifts. But a redshift of ~ 1100 *does* give a self-consistent result: the input and output redshifts agree to a precision of two significant figures, and so this is the best estimate for z_{ls} .

The relatively crude approach we have taken here has shown that decoupling takes place at a redshift of ~ 1100 – a result very close to the accepted value for the redshift of CMB production, as quoted in Chapter 6, of $z_{\text{ls}} \sim 1090$. The example has also shown that the physics of decoupling, and therefore the precise timing of the production of the CMB, depends strongly on how recombination proceeded. This brings us back once again to the crucial importance of the baryon-to-photon ratio (which helps determine the time of recombination) for cosmology.

9.2 Measuring primordial abundances

Elemental abundances are a fundamental tool in astronomy. All of the elements in the present-day Universe were either synthesised in the big bang or are produced by the evolution of stars.

In the previous chapter you saw that a small number of elements were synthesised in the very early Universe, before the temperature dropped below that needed for further fusion. You also saw that the fusion processes were highly sensitive to the baryon-to-photon ratio, η , and the neutron-to-proton ratio at early times. If the theory we have presented is correct, the start and end times for big bang nucleosynthesis – and so the yields for different elements – depend mainly on the value of η . Hence measuring elemental abundances in astrophysical environments provides a useful cosmological test, which is very complementary to other measurements, such as CMB anisotropies.

- Which cosmological density parameter is closely related to the baryon-to-photon ratio?
- The baryon density parameter, Ω_b , is closely related to η .

In this section we will explore how observations can be used to measure the abundances of the elements produced in the early Universe. We will also examine how these measurements can be used to confirm the theories of early-Universe physics presented in the last few chapters.

9.2.1 Abundances and metallicity

The majority of measurements of elemental abundances come from the use of **spectroscopy**, whereby the light from stars, gas or other sources of emission is measured as a function of its wavelength or frequency. This method allows us to observe emission or absorption lines caused by the presence of particular elements, as shown in Figure 9.2 for a relatively nearby galaxy.

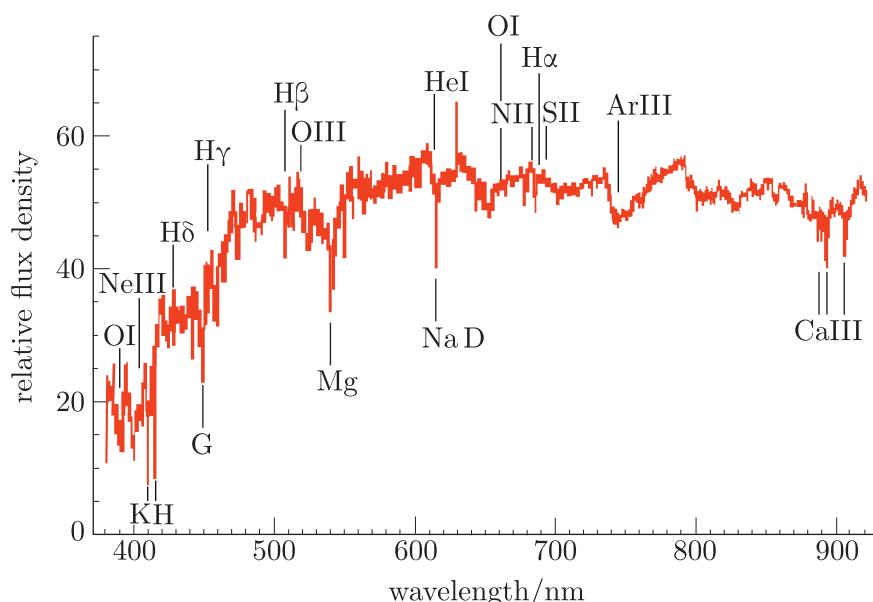


Figure 9.2 An optical spectrum of a galaxy, observed as part of the Sloan Digital Sky Survey. Several emission and absorption features are labelled.

Measuring the strength of spectral line features for galaxies and gas clouds enables the **relative abundance** of specific elements to be determined. For example, the abundance of oxygen relative to hydrogen is commonly used as a proxy for the total amount of **chemical enrichment** (the increase in proportion of elements heavier than helium) that has taken place due to the synthesis of new elements in stars.

Elemental abundances can be reported in several different ways. The simplest abundance measure is the mass of a given element relative to the mass of hydrogen within the same sample (e.g. a gas cloud from which a spectrum has been measured). In this book we will write abundance ratios of this sort as a ratio of the two species, so ${}^4\text{He}/\text{H}$ is the mass ratio of helium-4 relative to hydrogen.

It is also common to make use of abundances defined relative to those observed in the spectrum of the Sun because these are very well determined. This relationship is defined as

$$[\text{O/H}] = \log_{10} \left(\frac{n_{\text{O}}}{n_{\text{H}}} \right) - \log_{10} \left(\frac{n_{\text{O},\odot}}{n_{\text{H},\odot}} \right) \quad (9.10)$$

where $n_{\text{O}}/n_{\text{H}}$ is the ratio of the oxygen and hydrogen number densities measured in the same object of interest, and the right-hand log term is the same ratio measured for the Sun. Oxygen can be replaced by any other element in this expression. The square brackets indicate that we are describing a logarithmic abundance ratio by number density.

Abundance ratios of heavy elements, such as oxygen and iron, are related to a quantity known as **metallicity**, Z . This describes the proportion of baryonic matter (by mass) that is made up of elements other than hydrogen and helium. Such elements are referred to as ‘metals’ in the context of astronomy, and changing metallicity in stars and galaxies is used to track star formation throughout cosmic history. Metallicity is defined as:

$$Z = 1 - (X + Y) \quad (9.11)$$

where X and Y are the mass fractions of hydrogen and helium, respectively, for a given environment.

The abundance ratios $[\text{O/H}]$ and $[\text{Fe/H}]$ are often used to measure Z , so metallicity is sometimes used loosely to refer to one of these observed abundance ratios. A quantity such as $[\text{O/H}]$ can be related to Z by assuming that oxygen makes up a specific fraction of the metals, for example the same fraction that it does in the Sun.

The use of abundances and metallicity is widespread in extragalactic astronomy, where it is a powerful way to investigate the nuclear reactions that have historically taken place as a result of star formation in a particular region of the Universe (e.g. within the Milky Way). But we are not able to directly observe and measure spectral features for regions in the period immediately after the end of nucleosynthesis, and abundance measurements for a present-day galaxy may not reflect the primordial abundances created immediately after the big bang. So how can abundance measurements be used in cosmology?

The answer is that to infer early-Universe abundances, great care needs to be taken to choose the best locations to make measurements, and then considerable thought is needed to interpret them using all of the information available. In the next sections you will meet some specific examples of how this is done.

9.2.2 Deuterium abundances

Deuterium was the first element to be synthesised in the early Universe. Its abundance provides an important cosmological test, because its production was so sensitive to the relative proportion of energetic photons available to dissociate nuclei as they formed (i.e. η). Figure 9.3 shows a prediction for how the cosmic abundance of deuterium (D/H) depends on η .

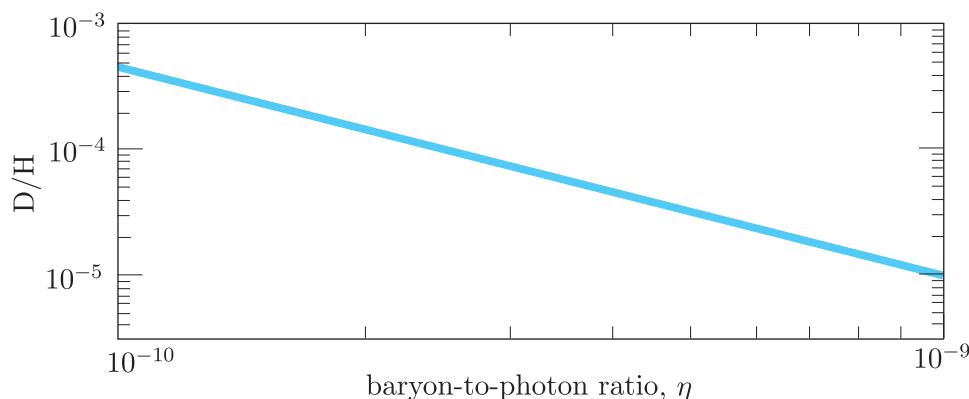


Figure 9.3 The dependence of deuterium abundance on η (after Cyburt *et al.*, 2016).

Deuterium is the primordial element whose abundance can be measured most reliably because it was *not* produced in significant quantities after the big bang nucleosynthesis (BBN) period. It can only decrease via **astration**, where it is destroyed via processes occurring inside stars. The quantity present in the Universe today must therefore be a lower limit on the amount that was produced during the BBN period.

Deuterium has been measured in several different astrophysical locations over many decades. The two main environments in which it is measured are the local interstellar medium (i.e. regions within the Milky Way) and the environments of very distant quasars. The latter are bright, distant galaxies whose central black hole is rapidly accreting gas, making them very luminous and so visible out to large distances. Absorption features in their spectra reveal the abundance of intervening intergalactic gas.

Figure 9.4, taken from a study by Cooke *et al.* (2014), plots a compilation of deuterium abundance measurements obtained from spectra of high-redshift quasars. It shows that the values along the paths travelled by light from different quasars cluster quite closely around an abundance value of $\log_{10}(D/H) = -4.6$. The coloured band indicates the predicted deuterium abundance based on the *Planck* measurement of $\Omega_{b,0}$.

Drawing conclusions about early-Universe abundance is harder for other elements, where both production and destruction may have taken place at later times.

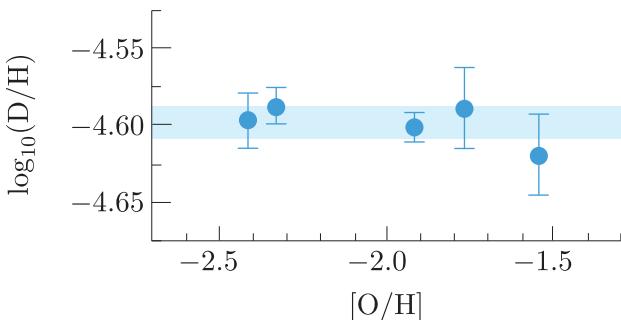


Figure 9.4 Observational measurements of deuterium abundance from the spectra of high-redshift quasars (blue data points) relative to the predicted abundance based on *Planck* measurements (blue band).

The next example considers another set of abundance measurements determined by observing objects within the Milky Way.

Example 9.3

Table 9.3 lists a set of D/H ratios measured from absorption features in the ultraviolet (UV) spectra of stars from different regions of the Milky Way.

Table 9.3 Deuterium abundance measurements from UV sightlines to selected Milky Way stars, with uncertainties.

Stellar sightline	D/H ($\times 10^{-6}$)
BD +28 4211	13.9 ± 1.0
BD +39 3226	11.7 ± 3.1
δ Ori	7.4 ± 1.0
Feige 110	21.4 ± 4.8
G191–B2B	16.6 ± 4.1
GD 246	15.1 ± 1.9
γ Cas	9.8 ± 2.5
HD 191877	7.8 ± 2.0
HD 195965	8.5 ± 1.6
HZ 43	16.6 ± 1.4
ι Ori	14.1 ± 2.8
LSS 1274	7.6 ± 1.9
TD1 32709	18.6 ± 5.3
WD 1034+001	21.4 ± 5.3
WD 1634–573	15.8 ± 2.5

(Adapted from Linsky *et al.*, 2006, p. 1118.)

Use these data, and the additional mathematical information provided below, to answer the following questions.

- Calculate the mean and median values of D/H.
- Next, determine the uncertainty on the mean.
- Finally, comment on how these measurements compare with the quasar absorption measurements shown in Figure 9.4, for which the mean value is $D/H = (2.52 \pm 0.02) \times 10^{-5}$.

Recall that the uncertainty in the mean value for a set of measurements is given by the **standard error on the mean**, σ_m , which is related to the **standard deviation**, σ , via:

$$\sigma_m = \frac{\sigma}{\sqrt{N}} \quad (9.12)$$

where N is the number of measurements.

The standard deviation, σ reflects the spread of values measured, and is defined as:

$$\sigma^2 = \frac{1}{N} \sum_i (x_i - \langle x \rangle)^2 \quad (9.13)$$

where x_i are the measured quantities and $\langle x \rangle$ is the mean of the x_i values.

Solution

One of the simplest ways to do these calculations is to write a short Python script that reads the list of numbers into an array, and then uses functions in the numpy library (e.g. `np.mean`, `np.median`, `np.std`) to do the calculation. However, we will work through the calculations by hand here.

- (a) Summing the values – ignoring uncertainties – and diving by the number of measurements gives a mean value of 1.38×10^{-5} . The median is 1.41×10^{-5} . The fact that the mean and median are very similar provides useful reassurance that the mean is a good approximation to the centre of the distribution of values.
- (b) To determine the uncertainty we must first calculate σ , and then σ_m . Calculating the standard deviation manually requires determining $x_i - \langle x \rangle$ for each row in Table 9.3 (where $\langle x \rangle = 1.38 \times 10^{-5}$ as calculated above), and then applying Equation 9.13. Whether you do this by hand or use `numpy.std`, you should obtain a value of $\sigma = 4.6 \times 10^{-6}$.

Since there are 15 measurements, this means that $\sigma_m = 1.2 \times 10^{-6}$.

- (c) We can now make a useful comparison with the quasar measurements in Figure 9.4. The mean value and uncertainty are given in the question as 2.52×10^{-5} and 2×10^{-7} , respectively.

Therefore, accounting for the uncertainty on the two mean values, the mean abundance estimate from the Milky Way (1.38×10^{-5}) is inconsistent with the quasar absorption measurements. The Milky Way deuterium content therefore appears to be significantly lower than that in the environments of high-redshift quasars.

- Which location – the Milky Way or the environments of quasars – is likely to contain gas whose composition better reflects the primordial Universe?

- Distant galaxies are being observed at an earlier time in the history of the Universe, because of the time taken for their light to reach us. It would therefore be expected that their environments will have been less affected by processes that alter the gas composition, since they will have operated over a much shorter time period.

Exercise 9.2

Use Figure 9.3 to find the values of η corresponding to the measured deuterium abundance determined via: (i) the high-redshift quasar measurements that you saw in Figure 9.4; (ii) the analysis of the Milky Way measurements in Example 9.3.

Calculate the corresponding value of $\Omega_{b,0}$ for each one, assuming the mean particle mass is m_p and the present-day photon number density is $n_{\gamma,0} = 4.0 \times 10^8 \text{ m}^{-3}$.

As the previous example and exercise show, the quasar measurements of D/H provide a tighter constraint than measurements of Milky Way regions, and they also result in a value of η that is in better agreement with the baryon density measured from the CMB. Since we already have grounds to expect the high-redshift measurements to be more reliable, the consistency between the D/H abundances and CMB measurements is a very impressive success of big bang theory. It seems likely that the Milky Way deuterium abundance has been depleted by processes occurring at more recent times.

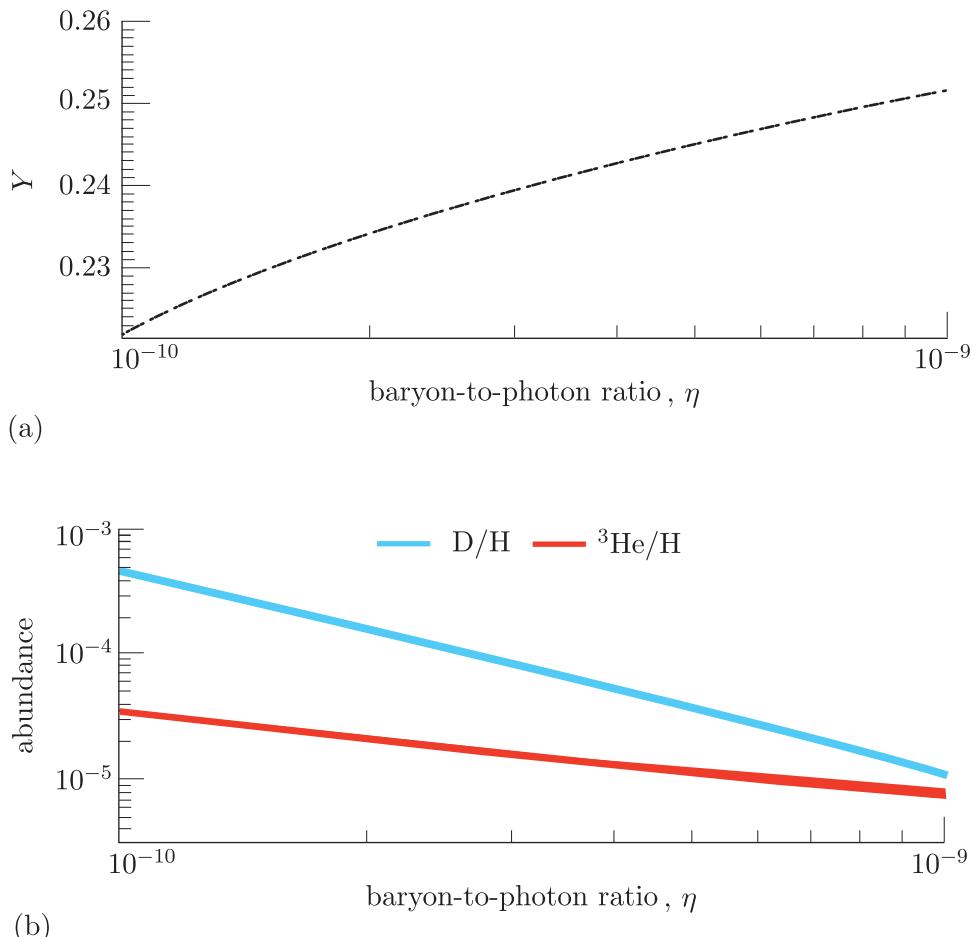
Another interesting route to measuring primordial abundances is to look within the Solar System. The abundance of deuterium in the atmospheres of Jupiter and Saturn has been measured to have a value of $D/H \sim 2 \times 10^{-5}$ (Lellouch *et al.*, 2001). Since no nuclear synthesis takes place in planetary atmospheres, these values should reflect the elemental abundances in the Universe at the time the Solar System formed, $\sim 4.5 \times 10^9$ years (or 4.5 gigayears, Gy) ago. The lookback time to galaxies at a redshift of ~ 3 (assuming *Planck*-measured cosmological parameters) is ~ 12 Gy, so the Solar System and quasar deuterium abundance measurements also appear consistent, assuming a fraction of the deuterium has been astrated in the time prior to Solar System formation.

9.2.3 Helium abundances

Helium fusion can only take place once the deuterium bottleneck is overcome, so the abundance of helium can also be used to measure η and Ω_b .

- Which two forms of helium are produced via BBN, and how do they differ?
- Both ${}^3\text{He}$ (helium-3) and ${}^4\text{He}$ (helium-4) are created.

Figure 9.5 shows the dependence of both of these helium isotopes on η , with D/H shown for comparison. Both ^4He (shown as Y) and D/H have a strong dependence on the baryon-to-photon ratio, whereas ^3He is less sensitive to it.



You may wonder why ^4He is expressed as a mass fraction, but the other isotopes as abundance ratios. Y is an intuitive quantity, as it tells us that $\sim 25\%$ of matter is in the form of ^4He . The mass fractions for deuterium and ^3He are so much lower that it makes more sense to think about ratios of numbers of particles.

Figure 9.5 The dependence of (a) 4-helium abundance (expressed as the mass fraction, Y) and (b) of 3-helium abundance on η , with deuterium abundance (D/H) shown for comparison (after Cyburt *et al.*, 2016).

The next exercise invites you to consider further the interpretation of Figure 9.5.

Exercise 9.3

Explain why the abundance of helium-4 is expected to *increase* for higher values of η , whereas the abundance of deuterium *decreases*, as shown in Figure 9.5.

Measuring helium abundances is therefore, in principle, an independent cosmological test that can provide different information from the

deuterium measurement. Helium abundance is also important for the conversion between η and $\Omega_{b,0}$, because this calculation depends on the mean particle mass at present times. It is harder to obtain reliable measurements of the primordial abundance for helium than for deuterium, because the importance of helium production within stars means that its abundance has changed a lot over the history of the Universe. However, it is still possible to achieve interesting and useful results that are helpful for cosmology by identifying regions to study where relatively little star formation has taken place.

- How can regions with a history of low star formation be identified?
- The metallicity of galaxies and extragalactic gas clouds traces the amount of star formation that has taken place. Low-metallicity regions can be identified by their spectra, and should correspond to locations where the gas has been least processed by stars.

Figure 9.6 shows a compilation of helium abundance measurements taken from low-metallicity extragalactic **HII regions** – these are regions of ionised hydrogen gas in distant galaxies.

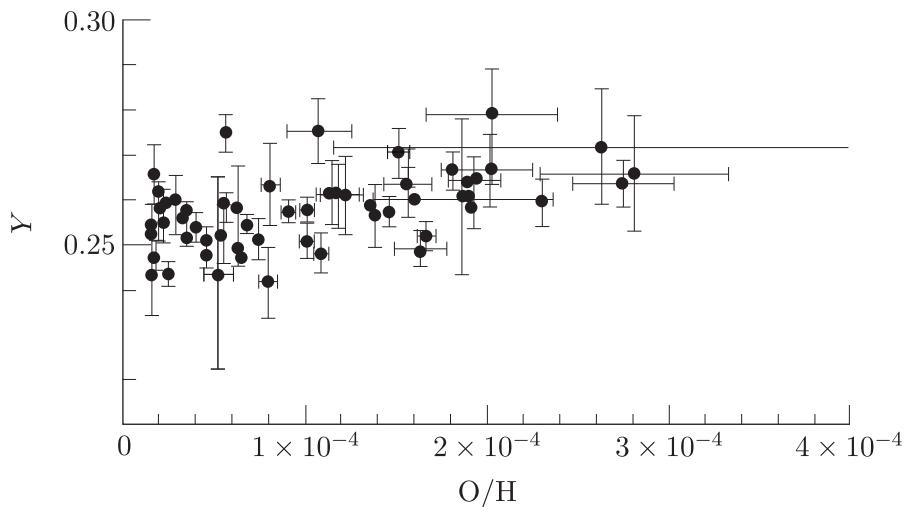


Figure 9.6 Measurements of helium abundance (Y) from extragalactic HII regions plotted against relative oxygen abundance for those regions, which is a measure of their metallicity.

You might notice that the amount of scatter in the helium abundances is much larger than for the quasar deuterium measurements in the previous section. This is not surprising given the effect of stellar processing on helium. However, it is possible to use the slight dependence of abundance on metallicity to extrapolate a value for a metallicity of zero, which would correspond to the primordial helium abundance. Extrapolation of these data suggests an Y abundance of ~ 0.25 at that time.

9.2.4 Lithium abundance

Measurements of cosmic lithium abundance have presented a challenge for cosmologists for a number of years. Lithium is the only element in the Universe that was produced in the big bang, but is also produced in stars and via interactions with cosmic rays within the interstellar medium. This makes it challenging to interpret observational measurements.

A number of attempts have been made to relate lithium abundance to metallicity, in order to be able to extrapolate back to a primordial value. There have been conflicting results on how this trend behaves, but one example from Sbordone *et al.* (2010) is shown in Figure 9.7.

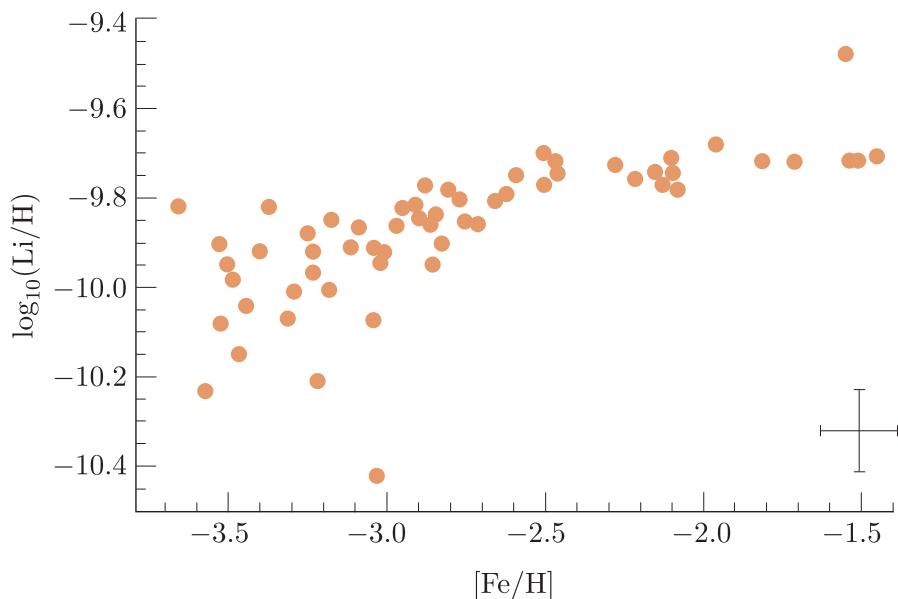


Figure 9.7 Observational measurements of lithium-7 abundance (plotted as a logarithmic abundance) compared to metallicity, in this case measured via iron abundance. The cross indicates the typical size of measurement uncertainties.

Unfortunately there is a lot of scatter in lithium abundance measurements at low metallicity compared with high metallicity. But a wide range of studies have reached the conclusion that – unlike the well-determined results for deuterium and helium – the primordial abundance of lithium-7 appears to be somewhat *lower* than that predicted by BBN, which is plotted in Figure 9.8.

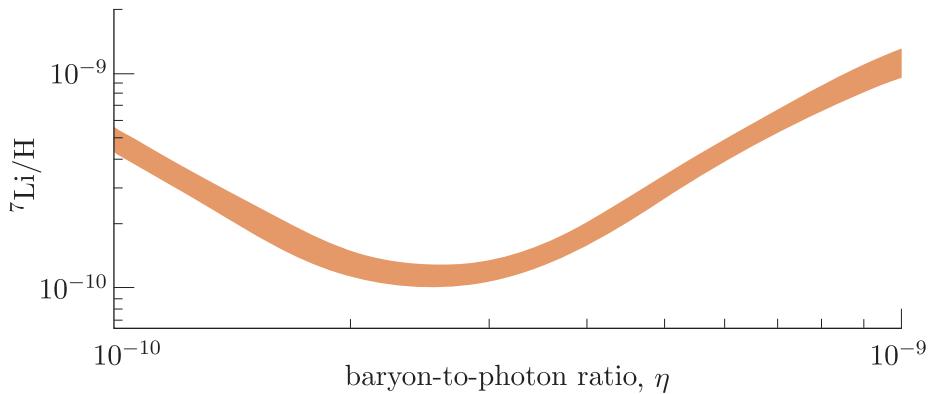


Figure 9.8 The dependence of lithium-7 abundance on η (after Cyburt *et al.*, 2016).

- What is the predicted value of the lithium-7 abundance for the *Planck*-measured value of $\eta = 6.1 \times 10^{-10}$, and how does this compare to the measurements in Figure 9.7?
- For $\eta = 6.1 \times 10^{-10}$, the predicted abundance inferred from the plot is $(^7\text{Li}/\text{H}) \approx 4.5 \times 10^{-10}$. To compare with the observed values shown in Figure 9.7 we need to take the logarithm to obtain $\log_{10}(^7\text{Li}/\text{H}) = -9.3$. The prediction is therefore considerably higher than the observed measurements.

The most widely accepted explanation for this result is that there is a process taking place in which stars can destroy lithium in their outer atmospheres. An alternative is that some aspect of early Universe physics affecting only the final stage of nucleosynthesis is not yet understood. Although this is a long-standing uncertainty, the potential explanations are not thought to suggest any major problem in our overall understanding of early Universe physics.

9.2.5 The cosmological significance of Ω_b

The previous sections have demonstrated that, although they can be challenging to interpret, observed measurements of elemental abundances provide important confirmation of our understanding of early Universe physics, and specifically the baryon-to-photon ratio. You have seen that observational determinations of $\Omega_{b,0}$ from elemental abundances conclude that its value is ~ 0.05 , with a similar value now resulting from CMB observations.

- What do the observational estimates of $\Omega_{b,0}$ imply for the influence of (ordinary) baryonic matter on the expansion of the Universe, if we assume that the Universe is spatially flat (i.e. $k = 0$)?
- A flat Universe has $\Omega = 1$, so if $\Omega_{b,0} \ll \Omega$ then this implies that other elements, such as dark matter and/or dark energy, must play the dominant role in controlling the expansion of spacetime in the present-day Universe.

Elemental abundance measurements have been very important in the historical development of cosmology, because they provide a key piece of evidence that there is not enough baryonic matter in the Universe to explain its rate of expansion. The next section therefore considers the need for non-baryonic matter.

9.3 Measuring non-baryonic matter

The primordial abundances of helium and deuterium discussed in the previous section are part of a range of important clues that point to the existence of some form of dark matter. We were able to neglect the role of dark matter in our investigation of how nuclei and atoms form, but you have already seen that it is part of the story of how the CMB originated, and may have noticed its inclusion in the pictorial summary of early-Universe evolution in Figure 8.5. Its role becomes especially important in the next part of the history of the Universe, in which matter gradually assembles to form stars and galaxies.

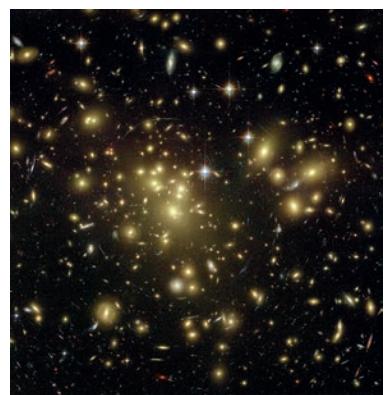
In this section we will consider the variety of evidence that has led to the current consensus that dark matter must be an important constituent of the Universe. We will start by examining how observations of nearby galaxies and galaxy clusters provided the first evidence for dark matter, and now provide very tight constraints on how any form of dark matter (or alternative theory) must behave. We will then return to the topic of the CMB and discuss how its anisotropies, together with modern theories of structure formation, point firmly to the existence of dark matter as an explanation for how galaxies assemble.

9.3.1 Weighing galaxy clusters

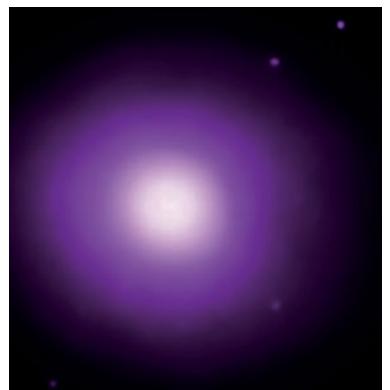
Galaxy clusters are groups of large numbers of galaxies that comprise the largest gravitationally bound structures in the Universe. Both their brightness and the fact that they contain a representative sample of the matter distribution throughout the Universe mean that they provide ideal ‘laboratories’ to understand the composition of the Universe.

Figure 9.9 shows two of the main constituents of galaxy clusters for the same object, Abell 1689. Panel (a) is an optical image of the individual galaxies that are members of this cluster, with all their constituent stars, whereas panel (b) is an X-ray image of a large quantity of hot, diffuse gas spread out in between the galaxies, which has been heated to X-ray-emitting temperatures through falling into the gravitational potential well of the cluster itself.

One of the first astronomers to estimate the total mass contained within a galaxy cluster was Fritz Zwicky, who, in 1933, measured the velocities of the galaxies in the Coma cluster. The galaxies in a cluster are expected to orbit the cluster’s centre of mass with velocities that depend on their mass (due to mutual gravitational attraction). However, Zwicky found that the



(a)



(b)

Figure 9.9 The galaxy cluster Abell 1689 viewed at (a) optical and (b) X-ray wavelengths.

range of velocities of the galaxies was considerably larger than expected. He used the **virial theorem** – a fundamental theorem, of wide use in astronomy – to calculate the total mass in the galaxy cluster and compared this with the mass contained in the individual galaxies and their stars.

Virial theorem

For an assembly of particles in stable equilibrium, the total gravitational potential energy (E_g) due to mutual attraction is related to the total kinetic energy (E_k) according to:

$$2E_k + E_g = 0 \quad (9.14)$$

In other words, the magnitude of the total kinetic energy is half that of the gravitational potential energy.

To apply the virial theorem to the comparison of total gravitational potential and kinetic energies in a galaxy cluster we need to be able to sum over particular properties of the cluster at different distances from the centre. This analysis requires the use of a volume integral, which you can read more about in the box that follows.

Volume integrals

It is common in astrophysics to encounter situations where we want to evaluate the sum of a quantity that varies over a volume or a surface. If we can model a star, a star cluster or a galaxy cluster as spherical, this means we can integrate that quantity over a range of radii, from $r = 0$ (the centre of the sphere) to $r = R$ (the outer radius).

Consider a gas cloud whose density ρ depends on radius, according to $\rho = \rho_0 [1 + (r/r_0)^{-2}]$, where r_0 is a constant representing a characteristic radius. To determine the total mass of that gas cloud we need to account for the variation of density with radius. We can therefore divide the sphere into a series of concentric spherical shells, each with its own density $\rho(r)$. The total mass of the gas cloud is the sum of all of the shells, which – if we make the shell width infinitesimally small – is given by:

$$M = \int_0^R \rho(r) dV$$

where dV is the volume of a shell, calculated as a spherical surface area $4\pi r^2$ multiplied by a shell thickness dr , so that:

$$M = \int_0^R 4\pi r^2 \rho(r) dr$$

Therefore, for the particular density distribution given,

$$M = \int_0^R 4\pi r^2 \rho_0 [1 + (r/r_0)^{-2}] dr = 4\pi \rho_0 r_0^2 \int_0^R [(r/r_0)^2 + 1] dr$$

which evaluates to $M = (4/3)\pi \rho_0 r_0^3 [(R/r_0)^3 + 3R/r_0]$.

We can use the virial theorem to estimate the total mass of a galaxy cluster, because both the gravitational potential energy and the kinetic energy depend on the mass. If m_i is the mass of an individual galaxy (i) within the cluster, and v_i is its time-averaged velocity, then the total time-averaged kinetic energy of the galaxy is given by:

$$E_k = \sum_i \frac{1}{2} m_i v_i^2$$

Of course the sum of all of the m_i values is the total mass, M , and so:

$$E_k = \frac{1}{2} M \langle v^2 \rangle$$

where $\langle v^2 \rangle$ is the average of the square of the velocities of the individual galaxies. We can't directly measure the overall galaxy velocities, but we can relate it to the **velocity dispersion**, σ_v , which is defined as the typical galaxy speed (magnitude of the velocity) in the radial direction. (We are able to use the Doppler shifting of emission lines in spectra to determine this component of velocity.) The velocity dispersion only measures the velocity in one dimension, but we can assume the distributions are similar in the other two dimensions, which means that $\langle v^2 \rangle \sim 3\sigma_v^2$. Therefore:

$$E_k = \frac{3}{2} M \sigma_v^2 \tag{9.15}$$

To obtain an expression for the total gravitational potential energy we need to consider the distribution of galaxies with radius in the cluster. The simplest assumption to make is that the galaxies are distributed uniformly within a spherical region of radius R , which means that we can define a galaxy density, ρ , that is independent of radius.

We can now use the volume integral approach explained above to sum the gravitational potential energy of a series of spherical shells, each of which spans a narrow range of radius from r to dr . The gravitational potential energy of a spherical shell at radius r is:

$$dE_g = -\frac{GM(r) dm}{r}$$

where $M(r)$ is the total mass contained at radii smaller than r and dm is the mass contained within the particular shell.

We can use the assumption of uniform density to find expressions for both $M(r)$ and dm . The mass within a spherical shell is then given by $4\pi r^2 dr \rho$, namely the volume of an infinitesimally thin spherical shell multiplied by the density of that region.

Substituting for $M(r)$ and dm and then integrating over all radii gives:

$$E_g = - \int_0^R \frac{G(4/3)\pi\rho r^3 4\pi r^2 \rho dr}{r} = -\frac{16}{3}\pi^2 G \rho^2 \int_0^R r^4 dr$$

which evaluates to:

$$E_g = -\frac{16\pi^2 G \rho^2 R^5}{15}$$

We require an expression in terms of the total mass M , which is $M(R)$, and so we can substitute back the relationship between ρ and M to find the expression for the total gravitational potential energy of a uniform density sphere:

$$E_g = -\frac{3GM^2}{5R} \quad (9.16)$$

We could now apply the virial theorem to relate our two total energies. However, we started with the simplifying assumption of uniform density, ρ . In reality, we know that galaxy clusters (like stars and galaxies) are centrally concentrated. This does not significantly change the analysis above, but the magnitude of the gravitational potential energy for a centrally concentrated mass distribution is higher by a small factor, so that it is common to neglect the factor of $3/5$ in Equation 9.16. In other words, the gravitational potential energy of a more centrally concentrated sphere is taken to be:

$$E_g = -\frac{GM^2}{R} \quad (9.17)$$

Now substituting our expressions for the two total energies into Equation 9.14 and rearranging for M gives a relationship between total mass, velocity dispersion and cluster radius:

$$M = \frac{3\sigma_v^2 R}{G} \quad (9.18)$$

Exercise 9.4

The Coma cluster has a velocity dispersion of $\sigma_v \sim 998 \text{ km s}^{-1}$, an estimated radius of $R \sim 3.5 \text{ Mpc}$ and an optical luminosity $L = 5.0 \times 10^{12} L_\odot$.

- Estimate the total mass, M , of the Coma cluster.
- Calculate the mass-to-light ratio, M/L , of the Coma cluster in units of solar mass divided by solar luminosity.
- Comment on what the calculations you've made tell us about the nature of the mass content in Coma. Could all of the mass be in the form of stars?

Exercise 9.4 demonstrates why Zwicky concluded that there must be something else present to explain the motions of the galaxies – clusters appear to have a significant quantity of ‘missing mass’.

Further corroboration of the need for dark matter in galaxy clusters comes from the X-ray-emitting hot-gas component shown in Figure 9.9b. The virial theorem does not just apply to galaxies: the hot gas should also obey it, with the kinetic energy of the gas related to the gravitational potential in the same way. Using Equation 9.16 and taking the total kinetic energy of a hot gas to be

$$E_k = \frac{3}{2} \frac{M}{\langle m \rangle} k_B T \quad (9.19)$$

where $\langle m \rangle$ is the mean mass of particles in the gas, the temperature of the intracluster gas should be related to the total mass of a galaxy cluster by:

$$T = \frac{GM\langle m \rangle}{5k_B R} \quad (9.20)$$

Exercise 9.5

X-ray spectroscopy of galaxy clusters shows that the temperature of the intracluster gas typically ranges from 10^7 – 10^8 K. Assuming a typical cluster radius of 1 Mpc and a typical particle mass of $\langle m \rangle = 0.6m_p$, calculate the range of total cluster mass (in units of solar mass) to which this corresponds.

You may wonder whether the mass of the X-ray-emitting hot gas, when added to that of the individual galaxies, could explain the missing mass. Unfortunately this is not the case. Very accurate estimates of the total gas mass in galaxy clusters can be made from X-ray observations, and although it makes up a larger proportion of the cluster mass than the stars within galaxies, the intracluster gas typically only accounts for around 10% of the total mass required by the virial theorem.

Perhaps one of the most compelling arguments for the existence of dark matter comes from multi-wavelength observations of a galaxy cluster known as the Bullet cluster. The Bullet cluster is a system in which two smaller groups of galaxies are merging together to form a single galaxy cluster. (This type of merger process is part of how present-day galaxy clusters assembled.)

Figure 9.10 shows the location of the two groups of galaxies (the circular and ellipse regions embedded in the blue regions) together with the location of the X-ray emission (shown in red). The emission forms two bright regions that are somewhat displaced from the groups of galaxies. The blue regions show the overall distribution of mass in the system, measured by a technique called gravitational lensing, in which the light from background galaxies is distorted as it passes near to the cluster.



Figure 9.10 Optical-light image of galaxies in the Bullet cluster. An overlaid X-ray image (red) and gravitational lensing map (blue) have been used to measure the hot gas and total mass distributions of the cluster, respectively.

The crucial point for interpreting these images is to understand why the X-ray emission is displaced. When galaxy clusters merge, the individual galaxies don't collide with each other, whereas the gas particles in the intracluster medium do. Hence the gas clouds are slowed down as they pass through each other, so that they have been 'left behind' by the galaxies.

If there was no dark matter then the total mass of the cluster would be dominated by the intracluster gas, which weighs much more than the galaxies. But the total mass is located in the same place as the galaxies, *and doesn't coincide with the gas*. This suggests the mass is dominated by dark matter, which – unlike the gas particles, but similarly to the galaxies – is expected to be 'collisionless', because dark matter particles interact only through gravity. Dark matter clumps would be expected to pass through each other and remain with the galaxies. There are no satisfactory explanations for the Bullet cluster observations that don't require the existence of dark matter.

9.3.2 Galaxy rotation curves

Zwicky's investigation of the motions of galaxies in clusters was the first observed evidence for the existence of some form of dark matter. But it was not until new evidence emerged in the 1970s from the rotation curves of spiral galaxies that the idea started to be taken more seriously within the astronomy community. The US astronomers Vera Rubin and Kent Ford made careful observations of the rotation speeds of stars at different locations within spiral galaxies. They found that – as for the orbits of entire galaxies within clusters – stellar orbits do not behave as theory would predict if only the visible stars and gas are present.

If stars are assumed to follow a circular orbit within the disc of the Milky Way, then their motion at a particular location should be described by **Keplerian rotation**, with the magnitude of the velocity given by:

$$v = \sqrt{\frac{GM(< r)}{r}} \quad (9.21)$$

where $M(< r)$ is the mass enclosed within a radius r .

- If most of the mass of a galaxy is concentrated close to the centre, how would you expect the velocity to change in the outer parts of the galaxy?
- If $M(< r)$ tends towards a constant value in the outer regions, then the velocity will decrease as r increases because of the inverse dependence on r .

Figure 9.11 shows an example of a spiral galaxy rotation curve, where the data points in panel (b) are the observed velocities. If there is no dark matter present, then the observed velocities should agree with those predicted from summing the mass of stars and gas (the dotted and dashed lines respectively, in the plot) according to Equation 9.21. This is clearly not the case: the velocity continues to increase in the outer regions.

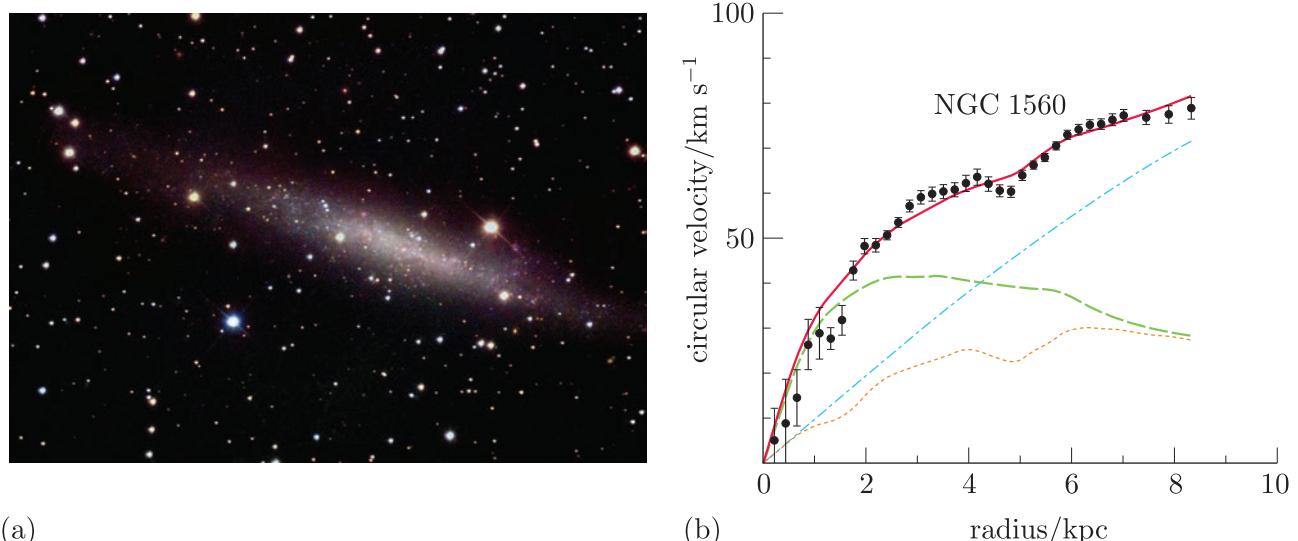


Figure 9.11 (a) An optical image of the approximately edge-on spiral galaxy NGC 1560 and (b) the galaxy's rotation curve. The green long-dashed curve shows the predicted contribution to the velocities from the inferred mass of stars, while the orange short-dashed line shows the contribution from the inferred mass of the gas. These are not enough on their own to explain the observed velocities. The blue dash-dotted line is the dark matter distribution required to make up the remaining mass.

The rotation curves of large samples of spiral galaxies are found to be either flat or increasing at radii where Keplerian theory suggests they should be decreasing. This has even been shown to be true of our own Galaxy. Figure 9.12 shows the Milky Way rotation curve, which is flat to distances of tens of kpc.

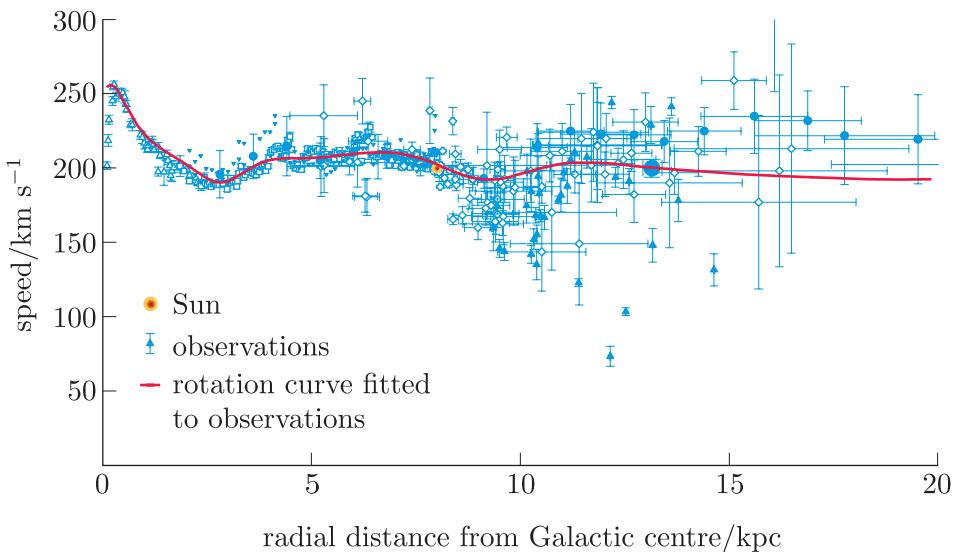


Figure 9.12 Circular velocity of the Milky Way plotted against radial distance from the galaxy centre (different symbols indicate several different sources of observational data).

The total mass of the Milky Way has been measured using the orbits of a large number of different types of tracers: stars, gas clouds, planetary nebulae, star clusters and small satellite galaxies at large distances. These measurements are generally in good agreement, and show that the Milky Way has a total mass of $\sim 10^{12} M_{\odot}$. This figure is around ten times the mass of the visible constituents of the Milky Way (stars, gas and dust), so it is believed that the Milky Way, like other galaxies, has a vast dark matter ‘halo’ that pervades the entire galaxy and extends to large distances.

9.3.3 Dark matter, the CMB and structure formation

In Chapter 7 you saw that CMB anisotropies are understood to originate in the oscillations of a fluid of photons and baryons in a ‘landscape’ of non-uniform density caused by perturbations in the initial distribution of matter. You also saw how the angular power spectrum can be used to measure Ω_b and Ω_m , which provides strong evidence for dark matter as the relative heights and locations of the power spectrum peaks require $\Omega_b \ll \Omega_m$.

But it is important to consider whether the entire framework we used to interpret the angular power spectrum could be incorrect. How would density perturbations evolve if there was *only* baryonic matter and radiation, with no dark matter valleys and hills in which to oscillate?

In fact, the overall uniformity of the CMB provides very strong evidence for the existence of dark matter. If there is no dark matter then all of the structure in the Universe – the galaxies and galaxies clusters we observe

today – must have grown from the gas density perturbations present at the time the CMB was emitted. The tiny amplitude of those observed fluctuations (~ 1 part in 10^5) means that the baryonic density perturbations at that time were too small to grow quickly enough via gravity to explain the structure that we observe today.

The model described in Chapter 7, of photon–baryon oscillations driven by the presence of perturbations in an underlying distribution of dark matter, allows for the presence of larger variations in density at early times: large enough to grow relatively quickly to form the cosmic web of structure (galaxies and galaxy clusters) that we see today.

All modern cosmological simulations start by modelling the growth of dark-matter perturbations. Structure formation theory also places an important requirement on the *nature* of dark matter: the majority of successful structure formation simulations assume **cold dark matter**. By ‘cold’, we mean that whatever type of unknown particle is involved must travel slowly, in contrast to ‘hot’ dark matter, which would be made up of relativistic particles. Cold dark matter is able to clump together, whereas other types are not expected to form stable structure so easily.

You will learn more about dark matter as the module continues. In the next two chapters we will finish exploring the history of the Universe by discussing how structure formation from the time of the CMB onwards led to the presence of stars and galaxies.

9.4 Summary of Chapter 9

- Recombination is the period in the early Universe when nuclei and electrons combined to form atoms. The time at which this happened depends on temperature and (less strongly) on the baryon-to-photon ratio, η .
- The process and timing of recombination influences when decoupling of matter and radiation takes place, because the main interaction process is Thomson scattering by free electrons, whose density becomes very low once atoms form at recombination.
- Decoupling occurs around the time of last scattering, so corresponds to the point at which the CMB is produced. The redshift of decoupling is related to the ionisation fraction X and cosmological parameters via:

$$z_{ls} = \left[\frac{H_0 \sqrt{\Omega_{m,0}}}{X(z) n_{b,0} \sigma_T c} \right]^{2/3} - 1 \quad (\text{Eqn 9.9})$$

and current best estimates give $z_{ls} \approx 1090$.

- **Elemental abundances** can be measured observationally, providing an important test for big bang nucleosynthesis (BBN) and for our understanding of early Universe physics. They provide a way to measure the cosmic baryon density, $\Omega_{b,0}$, because of the dependence of nucleosynthesis on η .

- Abundances are typically measured relative to the abundance of hydrogen, either as the relative proportion of a given element by mass or by particle number.
- Primordial abundances are difficult to measure, because many elements are also produced or destroyed in stars and galaxy environments. The best measurements come from regions of low **metallicity**, where the least processing by stars has taken place.
- Deuterium abundance has been measured most reliably via **spectroscopy** of high-redshift quasars. Their spectra contain absorption features caused by the presence of deuterium in the quasar's environment. The relative abundance of deuterium is found to be $D/H \sim 2.5 \times 10^{-5}$, which corresponds to a baryon density in good agreement with CMB measurements.
- Helium abundance can be measured from ionised gas regions outside the Milky Way: estimates from the lowest metallicity regions also show good agreement with BBN predictions.
- The inferred baryon density from cosmic abundances provides one strong piece of evidence for the existence of dark matter: there cannot be enough baryons present to explain the evolution of the Universe's expansion.
- The total gravitating mass of **galaxy clusters** can be measured in multiple ways, all of which demonstrate the need for dark matter.
- The **virial theorem** relates cluster mass to the **velocity dispersion** of the constituent galaxies:

$$M = \frac{3\sigma_v^2 R}{G} \quad (\text{Eqn 9.18})$$

while the hot gas temperature measured via X-rays also provides a mass estimate:

$$M = \frac{5k_B RT}{G\langle m \rangle}$$

- The rotation curves of spiral galaxies, including the Milky Way, also provide strong evidence for the existence of dark matter, because velocities of stars, gas clouds and satellite galaxies orbiting in their outer regions are higher than otherwise expected.
- Finally, the CMB anisotropies provide some of the strongest evidence for non-baryonic **cold dark matter**, because the perturbations we observe are too small to explain the growth of galaxies if only baryons were present, and because dark matter particles need to be relatively slow moving ('cold') to allow structure to form.

Chapter 10 Structure formation and the cosmic web

We have seen from the CMB that the early Universe had a very uniform density distribution, with only very small anisotropies (~ 1 part in 10^5). The CMB anisotropies visible today are thought to originate in quantum fluctuations in the distribution of matter and radiation that occurred at the time of inflation, which led to the acoustic oscillations of baryons and photons discussed in Chapter 7. The CMB anisotropies follow an angular power spectrum with distinct peaks that correspond to physical scales imprinted on the early Universe, and relate to the size of the horizon (the distance light can have travelled in the time so far elapsed) at that time.

In this chapter, and the one that follows, you will consider how the density perturbations present at the epoch of last scattering subsequently evolved to produce large-scale structure in the Universe. You will explore how gravity enabled the small density inhomogeneities of the early Universe to grow with time to produce galaxies, stars, planets and – ultimately – us.

Objectives

Working through this chapter will enable you to:

- explain the principles and key equations describing the gravitational collapse of matter to form stable dark-matter halos in virial equilibrium
- mathematically resolve tensions between local gravitational collapse and the overall expansion of the Universe
- describe qualitatively how realistic 3D collapse processes evolve and how they differ from simple spherical models
- discuss how numerical simulations are used to model structure formation and how they can be compared with observations.

10.1 Growth of density perturbations

10.1.1 Hubble flow

- Consider two galaxies isolated in space. Will they move towards or away from each other?
- The answer is that it depends on the circumstances. Gravity will try to pull them together; expansion of the Universe will tend to separate them. They will move apart unless they are close enough and massive enough for gravity to dominate.

Locally, we can observe the infall of galaxies towards us or their motion away from us by measuring their velocities. Galaxies falling towards us will

appear to be blueshifted (i.e. they have negative velocities); the light from galaxies that are moving away from us (i.e. with positive velocities) will appear redshifted.

The radial velocities of galaxies near to the Milky Way are shown in Figure 10.1, which can be compared with the Hubble diagrams in Figures 1.2 and 1.3. In Figure 10.1 we can identify two distinct regions. First, within about 1.3 Mpc, galaxies have random radial velocities with respect to us, which can be either positive or negative. These galaxies incorporate all of the Milky Way's satellites, as well as the Andromeda Galaxy (Messier 31) and its satellites (including Messier 32 and Messier 33). Any of these galaxies could merge with ours in the future; taken together, they define the Local Group of galaxies.

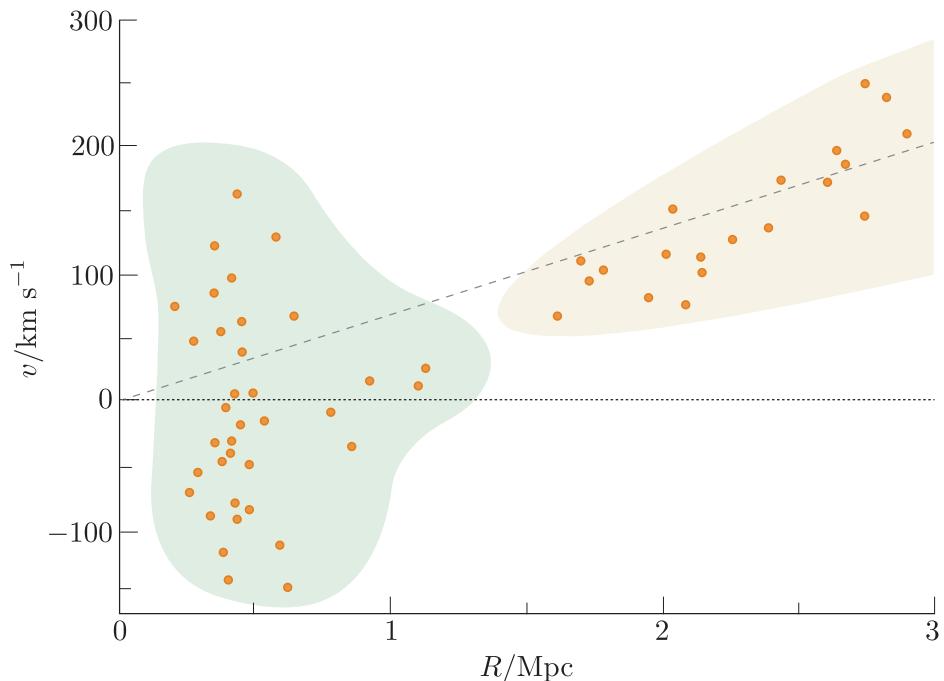


Figure 10.1 The observed relationship between distance and velocity for galaxies within 3 Mpc of Earth. The dashed line shows the Hubble–Lemaître relationship. The green region shows galaxies that appear bound to the Local Group; the orange-region galaxies have velocities that follow the Hubble flow.

The second region in Figure 10.1, outside the Local Group, shows galaxies whose radial velocity, V , is always positive, and becomes well correlated with distance from us, R . This is a sign that these objects are being carried away by the Universe's expansion. Although the galaxies in this region are still gravitationally attracted to us, the distance is too great and expansion too fast for them to ever merge with us, assuming the currently favoured model for the evolution of the scale factor, $a(t)$. These objects are described as being entrained in the Hubble flow.

Exercise 10.1

A galaxy lies 3 Mpc from the Milky Way and is receding from us at 200 km s^{-1} . Ignoring continued expansion of the Universe and the presence of any other galaxies, how long would it take gravity to slow the Milky Way and the other galaxy to rest? Assume that each has a mass of 10^{12} M_\odot .

10.1.2 Collapse or expansion?

We have previously discussed the interplay of gravitational attraction and gas pressure in the context of acoustic oscillations of the photon–baryon fluid at the time of CMB production (Chapter 7). Gravity draws matter (and radiation) inwards towards overdense regions, while gas pressure opposes this infall and causes matter to expand. This contest between gravity and pressure applies to many astrophysical situations – gravitational collapse of gas causes conversion of gravitational potential energy to thermal energy, increasing temperature and therefore pressure, and so potentially causing expansion. Stable stars form when these two forces are able to achieve balance. But in order for collapsed structures to form in the first place, the scales need initially to tip in favour of gravity.

We will start by ignoring dark matter, and considering the general case of a diffuse gas cloud, which, depending on its mass, might eventually form a galaxy or a cluster of stars. Such a cloud will be able to collapse if its mass exceeds a specific mass. This condition is known as the Jeans criterion.

The Jeans criterion

The **Jeans criterion** states that a cloud of gas will only collapse to form a stable, bound object if it is sufficiently massive for the gravitational attraction to exceed the initial pressure that supports it. The **Jeans mass**, M_J , is the minimum mass that a particular cloud of gas – of particle number density n and temperature T – must exceed in order to collapse:

$$M_J = \frac{9}{4} \times \sqrt{\frac{1}{2\pi n}} \times \frac{1}{\langle m \rangle^2} \times \left(\frac{k_B T}{G} \right)^{3/2} \quad (10.1)$$

where $\langle m \rangle$ is the average particle mass. This can also be written in terms of a **Jeans length**, λ_J , which is the size of a region of density ρ that contains the Jeans mass:

$$M_J = \frac{4\pi}{3} \rho \lambda_J^3 \quad (10.2)$$

where

$$\lambda_J \approx \sqrt{\frac{k_B T}{G \rho \langle m \rangle}}$$

In summary, a gas cloud of a given density and temperature needs to be more massive than M_J , or larger than λ_J , in order to collapse.

Another way to describe the situation is to ask whether an overdense region will grow with time if its density is greater than the mean background density, $\langle \rho \rangle$, by a factor δ , so that

$$\delta = \frac{\rho - \langle \rho \rangle}{\langle \rho \rangle} \quad (10.3)$$

In other words, will the region become even more overdense? For a baryonic gas (which is assumed to behave as an ideal gas) this would be true if the mass in the overdensity exceeds M_J .

The concept of Jeans mass is helpful to understand the factors influencing gravitational collapse, but there are several reasons why the formulae above, which assume a baryonic gas, are too simplified to describe the growth of structure in the early Universe. The next section considers what additional factors must be included for a more accurate model.

10.1.3 Limitations of the Jeans criterion

The roles played by radiation and dark matter are two of the most important complications in how ordinary matter collapses to form galaxies.

- Recalling the discussion of acoustic oscillations in Chapter 7, how would you expect radiation to affect the growth of baryon overdensities in the early Universe?
- Radiation dominated the energy density of the Universe at early times. From general relativity – and discussions in Chapter 7 – we expect the energy density of radiation to contribute to the gravitation driving collapse. Radiation also dominated the *pressure* of the photon–baryon fluid prior to recombination and decoupling, which would increase the opposing pressure compared to a baryon-only scenario.

You will see in later sections that the loss of radiation pressure when the photon–baryon fluid decoupled at the time the CMB was produced is what enabled galaxy-mass overdensities of baryons to begin to collapse rather than oscillate. After the period of radiation domination, which ended after nucleosynthesis took place, and prior to recombination, the dominant form of matter according to the favoured cosmological model was dark matter.

- How would you expect the dark-matter content of the Universe to affect the growth of baryon overdensities in the early Universe?
- Dark matter is assumed to dominate the matter content of the Universe at all times. As discussed in Chapter 7, around the time of decoupling, baryons (together with photons) oscillated within dark-matter overdensities. Therefore, once matter dominates radiation, the collapse of ordinary matter to form galaxies needs to account for the dominant gravitational influence of dark matter.

The Jeans analysis cannot be applied directly to dark-matter collapse. Particle dark matter does not interact electromagnetically. Collapsing dark matter can be considered to have an effective pressure, related to the increasing kinetic energy of the dark-matter particles, but its pressure cannot be defined in the same way as for an ordinary gas.

A final important consideration is that gravitational collapse takes place in an expanding universe, which changes the density and pressure of all of its components. Collapse to form a stable, bound object must happen on a timescale that is faster than expansion can separate the collapsing material.

10.2 Collapse of dark-matter halos

The remainder of this chapter looks at how dark-matter overdensities collapse, beginning with a uniform sphere and gradually making it more realistic. The next chapter then discusses how the baryonic gas later collapsed within these dark-matter overdensities to form galaxies. This order of discussion traces the process as it evolved through time, and also charts how simulations of the early Universe have had to evolve, to take into account the progressively more detailed physics of the real Universe.

Online resources: Python demonstrations

A set of Python demonstrations and short exercises based on solving the equations presented in this section is provided in the online module resources. These resources are designed to be relatively simple tasks that reinforce your understanding of these topics, so we strongly encourage you to complete these as you work through this section.

This will let you see how changing the initial conditions of the Universe would change the structure that evolves.

10.2.1 Collapse of a spherical overdensity

We start by modelling a spherically symmetric cloud of material, which can be modelled as a set of concentric spherical shells of matter. (See Section 9.3.1 for discussion of a similar geometry of concentric shells.) The collapse of such a cloud is dictated by the shell theorems of Newton, briefly mentioned when deriving the Friedmann equations in Section 4.1.1, which describe the gravitational forces acting on a given spherical shell of matter.

Newton's shell theorems for a spherically symmetric body

- A body located at a radius smaller than that of a shell of matter experiences no net gravitational acceleration from that shell.
- A body outside a shell of matter experiences gravitational acceleration of the same magnitude and direction as if all the matter in that shell were concentrated at its centre.

Let's first simplify the Universe into a uniform medium of density $\langle \rho \rangle$, comprised entirely of pressureless matter (e.g. dark matter or diffuse baryons). We place into this Universe a single overdensity of $\rho = \langle \rho \rangle(1 + \delta)$ and radius R , and split this sphere up into infinitesimally thin shells.

Because we are performing this calculation in only one dimension, each shell can also be considered to represent the radial trajectory of an individual dark-matter particle. Take a moment to think how this sphere might collapse: which parts of the sphere will collapse in which order?

Newton's shell theorems mean we can treat the overdensity as equivalent to a point mass at radius R and ignore anything outside it. The overdensity has mass

$$M = \frac{4\pi R^3}{3} \langle \rho \rangle (1 + \delta) \quad (10.4)$$

Newtonian gravity dictates that a particle will accelerate towards the centre of the overdensity at a rate of

$$\begin{aligned} g &= \ddot{R} \\ &= -\frac{GM}{R^2} \\ &= -\frac{G}{R^2} \left[\frac{4\pi R^3}{3} \langle \rho \rangle (1 + \delta) \right] \\ &= -\frac{4\pi G}{3} R \langle \rho \rangle (1 + \delta) \end{aligned} \quad (10.5)$$

By modelling the evolution of R over time (or by starting from Kepler's third law), it can be shown that the time it takes the overdensity to collapse to a central point is given by the **free-fall timescale**:

$$t_{\text{ff}} = \sqrt{\frac{3\pi}{32G\langle \rho \rangle(1 + \delta)}} \quad (10.6)$$

Converting back from $\langle \rho \rangle(1 + \delta)$ to ρ , and evaluating the numerical constants, leads to the simpler expression

$$t_{\text{ff}} \approx 18.45 \text{ hours} \left(\frac{\rho}{1 \text{ kg m}^{-3}} \right)^{-1/2} \quad (10.7)$$

Exercise 10.2

Use information in Table 10.1 to calculate (i) the mean total density and (ii) the free-fall timescale for both the Milky Way and the Local Group.

Table 10.1 Masses and radii for the Milky Way and the Local Group.

System	Mass/ M_\odot	Radius/kpc
Milky Way	1.5×10^{12}	26
Local Group	3.7×10^{12}	~ 1000

- Note that Equation 10.6 depends only on the matter density. How quickly will shells near the centre of our overdensity collapse compared to those further out?
- Because we have been considering a homogeneous overdensity, all shells have the same density and so their free-fall timescales are identical. All particles within the sphere will reach $R = 0$ at the same time.

When $R = 0$, the entire sphere has collapsed to a point. The entire mass M is contained in this point, giving it effectively an infinite density ($\delta \rightarrow \infty$). If we were considering collapse of a baryonic gas, then this could be prevented by the fact that collapse would heat up the gas, leading to increasing pressure, which would create a pressure gradient and (at some point) halt the collapse. However, the situation for non-interacting dark matter is different. We therefore need a way to modify our model dark-matter Universe to stop it collapsing into (very massive) black holes.

Online resources: free-fall collapse demo

The online module resources include a Jupyter Notebook that models free-fall collapse by directly modelling how particles evolve under Newtonian gravity. Use this notebook to explore free-fall collapse and confirm the timescale calculations made in Exercise 10.2.

10.2.2 Incorporating expansion of the Universe

One simple modification we can make to our model is to include the expansion of the Universe. As you saw in earlier chapters, this expansion is usually characterised by the behaviour of the scale factor, $a(t)$, which evolves according to the overall matter and energy density. We will explore how expansion alters gravitational collapse by considering a spatially flat ($k = 0$), matter-dominated ($\Lambda = 0$) universe, which is an appropriate model for the early stages of gravitational collapse.*

We can take Equation 10.5, divide through by R , and write it as the sum of two terms:

$$\frac{\ddot{R}}{R} = -\frac{4\pi G}{3}\langle\rho\rangle - \frac{4\pi G}{3}\langle\rho\rangle\delta \quad (10.8)$$

Because $\langle\rho\rangle$ is the mean density of the Universe (at the time being considered), the first term represents the self-gravity of any typical region of the Universe of radius R . In our $\Lambda = 0$ model, this corresponds to the global deceleration of the Universe's expansion owing to the self-gravity of the matter within it, which occurs via changes to the scale factor, a . If we set $\delta = 0$ we obtain

$$\frac{\ddot{R}}{R} = -\frac{4\pi G}{3}\langle\rho\rangle \quad (10.9)$$

*The derivation in this section follows Ryden (2017), pp. 213–216.

If we consider a region of size $R = a$, then the left-hand side becomes \ddot{a}/a , and Equation 10.9 is equivalent to the cosmological acceleration equation (Equation 4.35) for the $k = 0$ matter-only ($\Lambda = 0$) case.

The second term in Equation 10.8 represents the local, additional self-gravity of the particular overdensity being considered. To compute this, we can first recognise that the mass of the overdensity is constant, so following Equation 10.4,

$$R^3 \propto \frac{1}{\langle \rho \rangle (1 + \delta)}$$

However, because $\langle \rho \rangle \propto a^{-3}$, this becomes

$$R \propto a(1 + \delta)^{-1/3} \quad (10.10)$$

We are interested in the rate of change of δ with time, because this describes how the overdensity grows. Taking the second time derivative of Equation 10.10 and using the chain rule for derivatives of $(1 + \delta)$ gives, after some manipulation:

$$\frac{\ddot{R}}{R} = \frac{\ddot{a}}{a} - \frac{1}{3}\ddot{\delta}(1 + \delta)^{-1} - \frac{2}{3}\frac{\dot{a}}{a}\dot{\delta}(1 + \delta)^{-1} + \frac{4}{9}\dot{\delta}^2(1 + \delta)^{-2} \quad (10.11)$$

When $|\delta| \ll 1$, this can be simplified[†] to

$$\frac{\ddot{R}}{R} \approx \frac{\ddot{a}}{a} - \frac{1}{3}\ddot{\delta} - \frac{2}{3}\frac{\dot{a}}{a}\dot{\delta} \quad (10.12)$$

We can now substitute for the left-hand side using Equation 10.8 to find

$$-\frac{4\pi G}{3}\langle \rho \rangle - \frac{4\pi G}{3}\langle \rho \rangle \delta \approx \frac{\ddot{a}}{a} - \frac{1}{3}\ddot{\delta} - \frac{2}{3}\frac{\dot{a}}{a}\dot{\delta} \quad (10.13)$$

We can relate the first term on the left-hand side to \ddot{a}/a via Equation 10.9 where a region of size a is considered. This cancels the first term on each side to leave

$$-\frac{1}{3}\ddot{\delta} - \frac{2}{3}\frac{\dot{a}}{a}\dot{\delta} = -\frac{4\pi G}{3}\langle \rho \rangle \delta \quad (10.14)$$

Remembering that the Hubble parameter $H = \dot{a}/a$, and rearranging slightly gives

$$\ddot{\delta} + 2H\dot{\delta} = 4\pi G\langle \rho \rangle \delta \quad (10.15)$$

We now have what looks like a relatively straightforward differential equation that describes how a density perturbation changes with time. In order to solve this for $\delta(t)$ and find out how the perturbation grows or shrinks, we need to recall that both H and $\langle \rho \rangle$ change with time as well.

We noted earlier that we are considering a matter-dominated, $\Lambda = 0$ Universe, which is an appropriate model for the early stages of structure

[†]This is because terms in $1 + \delta$ reduce to 1, and the final $\dot{\delta}^2$ term can be neglected because the square of a quantity $\ll 1$ must be much smaller than terms involving the unsquared quantity.

formation. In Chapter 4 you saw that in this model $a(t) \propto t^{2/3}$, which means that $H(t) \propto t^{-1}$, and $\langle \rho \rangle = \rho_m \propto t^{-2}$.

The following example uses this information to consider the possible solutions to Equation 10.15.

Example 10.1

With the appropriate proportionality constants (from the Friedmann equation), the proportionalities for H and $\langle \rho \rangle$ can be written in full as

$$H(t) = \frac{2}{3}t^{-1} \quad \text{and} \quad \langle \rho \rangle = \rho_m(t) = \frac{3}{8\pi G} H(t)^2 = \frac{3}{8\pi G} \frac{4}{9} t^{-2}$$

Use this information to show that

$$\delta(t) = k_1 t^{2/3}$$

where k_1 is a constant, is a valid solution to Equation 10.15, and therefore that small perturbations are expected to grow in size in the early matter-dominated Universe, even when expansion of the Universe is accounted for.

Solution

To show that the provided expression is a valid solution to Equation 10.15, we start by differentiating it (twice) to find the following expressions for $\dot{\delta}$ and $\ddot{\delta}$:

$$\dot{\delta} = \frac{2}{3}k_1 t^{-1/3} \quad \text{and} \quad \ddot{\delta} = -\frac{2}{9}k_1 t^{-4/3}$$

We now need to substitute these expressions into Equation 10.15, along with the expressions provided for the time-dependence of H and $\langle \rho \rangle$.

Substituting all four expressions in gives a left-hand side of

$$-\frac{2}{9}k_1 t^{-4/3} + 2\frac{2}{3}t^{-1}\frac{2}{3}k_1 t^{-1/3} = \frac{6}{9}k_1 t^{-4/3}$$

and a right-hand side of

$$4\pi G \frac{3}{8\pi G} \frac{4}{9} t^{-2} k_1 t^{2/3} = \frac{6}{9}k_1 t^{-4/3}$$

Therefore the two sides of the equation match, and so we have shown that $\delta(t) = k_1 t^{2/3}$ is a valid solution for the evolution of δ with time. Because the exponent is positive, this tells us that – in the early, matter-dominated situation being considered – overdense regions will grow with time.

Example 10.1 has shown that structure formation can proceed successfully in the early, matter-dominated Universe, despite the counteracting effect of the expansion of spacetime. We can also describe this growth as a function of the scale factor and therefore of the observable quantity (redshift) as follows:

$$\delta(t) \propto t^{2/3} \propto a(t) \propto \frac{1}{1+z} \tag{10.16}$$

It is important, however, to remember that our analysis assumed that $\delta \ll 1$, and so this relation becomes increasingly inaccurate as $\delta \rightarrow 1$.

The next exercise considers the implications for what we observe in the real Universe.

Exercise 10.3

The CMB anisotropies (e.g. Figure 6.9) correspond to density deviations of $\delta \approx 10^{-5}$ at the time of decoupling ($z \approx 1090$). At the time of writing, the galaxies with the highest redshift that have been reliably observed have $z \approx 13$. Use Equation 10.16 to show that the overdensities implied by the CMB could not have evolved into galaxies at $z \approx 13$. Based on what you read in earlier chapters, how might this inconsistency be resolved?

In this section you have seen how the expansion of the Universe can be incorporated into a basic model of gravitational collapse. However, this has not solved the problem of how dark matter is prevented from collapsing to an infinitely dense point. In order to address this, we need to think further about the nature of dark matter and how it must behave during gravitational collapse.

10.2.3 Collapsing dark matter and baryons

Models of the Universe with overdensities based on the 1 : 100 000 fractional temperature fluctuations of the CMB do not accurately reproduce observed galaxies: we require overdensities at the time of recombination to be at least 100 times greater. This tells us an important fact about the nature of dark matter.

The density of baryonic matter must be strongly coupled to the density of radiation at the time of recombination, so the overdensities cannot be enhanced using any kind of known, baryonic material. Instead, we need some kind of gas- or fluid-like substance that has mass, and so can create its own overdensities, but interacts weakly with the radiation and matter around it. This requirement strongly suggests that dark matter comprises some sort of particle beyond the Standard Model of particle physics, one that does not interact electromagnetically with radiation and ordinary matter particles.

Figure 10.2 illustrates the growth of density perturbations arising from dark matter, and the differing behaviours of baryons and radiation.

Panel (a) shows that the dark-matter overdensity, δ_{DM} , grows smoothly with time, a process that starts at around the time of matter–radiation equality ($z \approx 3400$). Panel (b) shows a more schematic view of the differing behaviours of dark matter (blue regions) and baryonic gas (white regions).

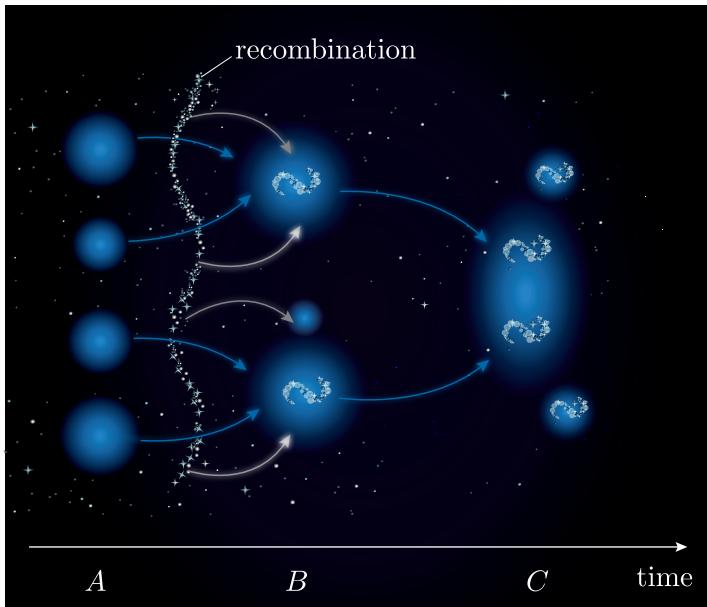
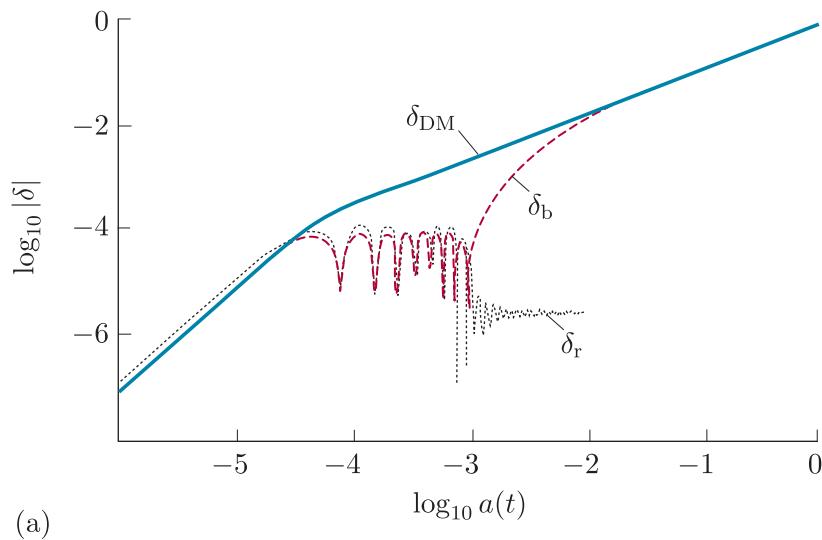


Figure 10.2 (a) The growth of density enhancements in radiation (δ_r), dark matter (δ_{DM}) and baryonic matter (δ_b); (b) a graphical illustration of this effect at three different times, A , B and C .

At time A , before recombination, overdensities of dark matter develop. These overdensities start clumping together into larger structures (blue arrows). At time B , after recombination, baryonic gas (white arrows) is decoupled from radiation, and can start pooling in these overdensities, forming the first galaxies. By time C , the overdensities have merged into largely isolated galaxy clusters, each containing many galaxies, which are themselves in the process of merging together.

10.2.4 Virialisation

In order to solve the problem of dark-matter collapse to infinite density, we need to assume a more realistic overdensity in which density increases smoothly towards the centre of the overdensity. In this situation the central, high-density parts have a shorter t_{ff} so collapse first; the outer parts take longer to collapse.

This has the consequence that smaller objects (with higher mean density) collapse faster than bigger ones; for example, halos on the scale of dwarf galaxies should collapse first, then begin to coalesce into larger objects like the dark-matter halos that host spiral and elliptical galaxies today.

If we consider a single large overdensity, with shells at different radii, then we can revisit how the collapsing shells evolve. Although each shell passes through $R = 0$, thereby theoretically creating infinite density, each infinitesimally thin shell has infinitesimal mass.

- What happens to a shell's energy as it passes through $R = 0$?
- Gravitational potential energy has been converted into kinetic energy, so the shell continues to expand again. However, the Universe has continued to expand and dilute around it, so instead it remains an overdensity, held together by gravity: it has become a **bound halo**.

Shells rebound at different times and so they will cross one another. This means that each shell no longer has the same mass inside it, and the gravitational potential at each radius changes rapidly. The repeated crossings of shells eventually stabilise to a constant, centrally concentrated mass profile. In the resulting stable configuration, dark matter particles at each radius are sustained by the balance of gravity and kinetic energy and obey the virial theorem (see Section 9.3.1).

It's worth considering how this kinetic energy manifests itself. An initially homogeneous thin shell of particles has become thoroughly mixed into a much more dynamic system of particles repeatedly passing one another and moving with respect to the centre of mass – in other words, behaving like a typical gas. Dark matter appears to be collisionless and non-baryonic, so we *cannot* equate it to a monatomic ideal gas with $\frac{1}{2}mv^2 = \frac{3}{2}k_B T$. Nevertheless, we can think of the distribution of kinetic energies in dark matter as being semi-equivalent to a kinetic temperature. In other words, it is the kinetic temperature of the dark matter that prevents its further collapse. Structure formation places requirements on this microphysical behaviour of dark matter, which helps to constrain models for candidate dark-matter particles. The collapse of baryonic gas within these dark-matter halos will be considered in the next chapter.

10.3 Collapse in three dimensions

10.3.1 Collapse of an ellipsoid

So far, we have only considered spherical collapse; however, it is more realistic to consider the collapse of a non-spherical object.

Online resources: ellipsoidal collapse demo

Another demonstration found in the online module resources leads you through the collapse and virialisation of a 3D ellipsoid, so you can visualise how this works in a more realistic setting.

Let's take our sphere and squash two of its dimensions until it becomes a cigar-shaped ellipsoid, as in Figure 10.3. Each axis of the ellipsoid will now collapse and virialise at a different rate, starting with the smallest. When the first axis collapses to $R = 0$, the ellipsoid turns into a flat ellipse, known as a **Zeldovich pancake**, after the physicist Yakov Zeldovich who first developed this theory. Collapsing and virialising the second axis forms a long, cigar-like **filament**; collapsing and virialising the third axis allows matter to flow along the filament into a central dark-matter halo.

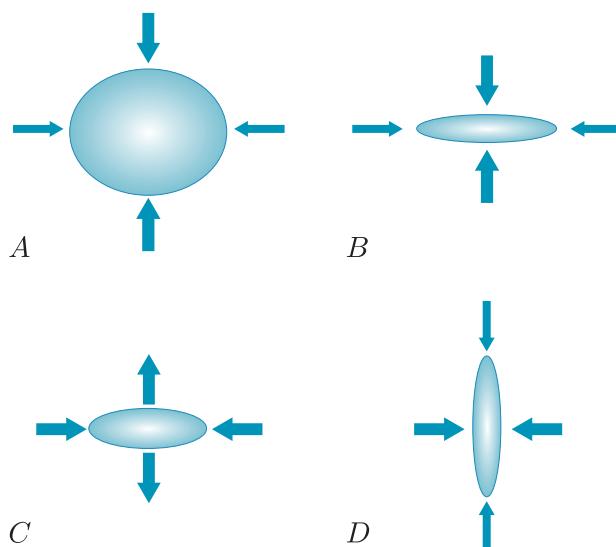


Figure 10.3 The collapse of an ellipsoid structure into a Zeldovich pancake, shown over four snapshots starting from time *A* to time *D*.

The slight ellipticity at time *A* means that the force along the shortest axis is stronger, so this axis collapses first, forming a flattened structure (time *B*). Remember that compression also works in the third dimension, which is not easily shown here. The vertical axis rebounds (time *C*) while the horizontal axis is still collapsing. Finally, the horizontal axis collapses (time *D*), by which point the vertical axis may already have rebounded to

its maximum extent. Note that the enhanced elongation (i.e. the pancake shape in three dimensions) is generally retained after the initial collapse.

Now let's consider many overlapping ellipsoids with voids between them. Matter will be pulled out of the voids into these collapsing ellipsoids, which will flatten into sheets (Zeldovich pancakes) joined together by filaments, at the junctions of which will be individual dark-matter halos. We're left with the foam-like structure of the cosmic web, as observed in the real Universe (e.g. as shown in Figure 1.1).

10.3.2 Numerical simulations

The real Universe follows the principles discussed in the previous sections, but is more complex than can be fully captured by this kind of intuitive reasoning and cartoon diagrams. Instead, it is necessary to turn to computer simulations like those of the online demonstrations for this chapter, but considerably more complex.

Online resources: numerical simulation videos

Visualising the complexities of structure formation in the Universe is difficult. The online module resources include links to videos of numerical simulations, where you can watch these collapse processes happen in physically realistic scenarios.

Cosmological simulations start with a near-homogeneous universe, populated with small overdensities and underdensities. The statistical properties of these density variations and their initial expanding motions are generated from an underlying model of cosmology, containing parameters calibrated against our observations of our Universe. Physics is encoded into the model to let the different components of the Universe (radiation, dark matter, baryonic matter and the cosmological constant) interact with themselves and each other according to the laws of physics.

Given these initial physical conditions, the simulation is begun and the model universe is stepped forward in small time steps. The gravitational physics incorporated into the model acts to enhance the initial overdensities so that they grow into the dark-matter halos in which galaxies and galaxy clusters form.

Types of simulation

Simulations must be programmed efficiently so that they run in the shortest time possible. This allows the Universe to be simulated in the smallest possible time steps, and with the highest possible spatial resolution, to ensure that physics happening on short timescales and in small regions is not ignored.

Astrophysical simulations tend to use one of a number of established modelling approaches. The choice of method depends on the physical

situation the programmers want to study. For example, this chapter discusses the collapse of dark-matter halos on large scales, and studies focusing on this collapse would typically use a different approach from the detailed studies of galaxy evolution discussed in the next chapter.

Commonly used methods include the following.

- ***N*-body simulations:** These split the Universe up into point masses, each with its own properties (e.g. mass). This makes them useful for looking at situations where objects behave mostly like particles, such as tracking the movements of star particles in galaxy collisions (note that these are not individual stars, but particles that each represent many stars). However, they are not very useful for simulating diffuse material, such as gas or dark matter. The Python demonstrations in this chapter are simple *N*-body simulations.
- **Smoothed-particle hydrodynamic (SPH) simulations:** These are essentially *N*-body simulations in which each particle is ‘smoothed’, meaning it covers a finite volume. These are useful in situations where there is complex motion, such as in the turbulent flows of galaxy and star formation.
- **Grid simulations:** These take the opposite approach to SPHs, by splitting the Universe up into cubes on a regular 3D grid and applying hydrodynamical equations to evolve the fluid within each cell, with grid cells assigned properties such as density, temperature, etc. Grid simulations are useful for the treatment of gas physics, but they are very inefficient if most of the simulation is empty space, and they can miss important details if the size of the grid is too large. Some grid simulations solve this with **nested sampling**, in which important grid squares are sampled by a sub-grid and can be incremented in smaller time steps than in the rest of the simulation. Grid simulations are also less useful if material moves around within the simulation, because they then have to track what goes through each face of each grid cube.
- **Moving-mesh simulations:** Moving or dynamic meshes solve the problems of empty space and moving material by splitting the Universe up into less-regular regions, tied together by a set of mesh points. These mesh points can be concentrated in the most dynamic regions of space and move with the material they are tracing. However, these still have limitations, for example difficulty in allowing media, such as stars and gas, to pass through each other, because material from both fluids is physically bound to the same mesh nodes.

Online resources: *N*-body simulation demo

A final demonstration for this chapter, also available in the online module resources, leads you through the first steps of building an *N*-body simulation. This basic set-up can be modified to simulate real systems in the Universe.

Figure 10.4 shows a set of snapshots from a sophisticated moving-mesh cosmological simulation called Illustris. The four rows show the evolution over four redshifts of dark-matter density, baryonic gas density, temperature and metallicity.

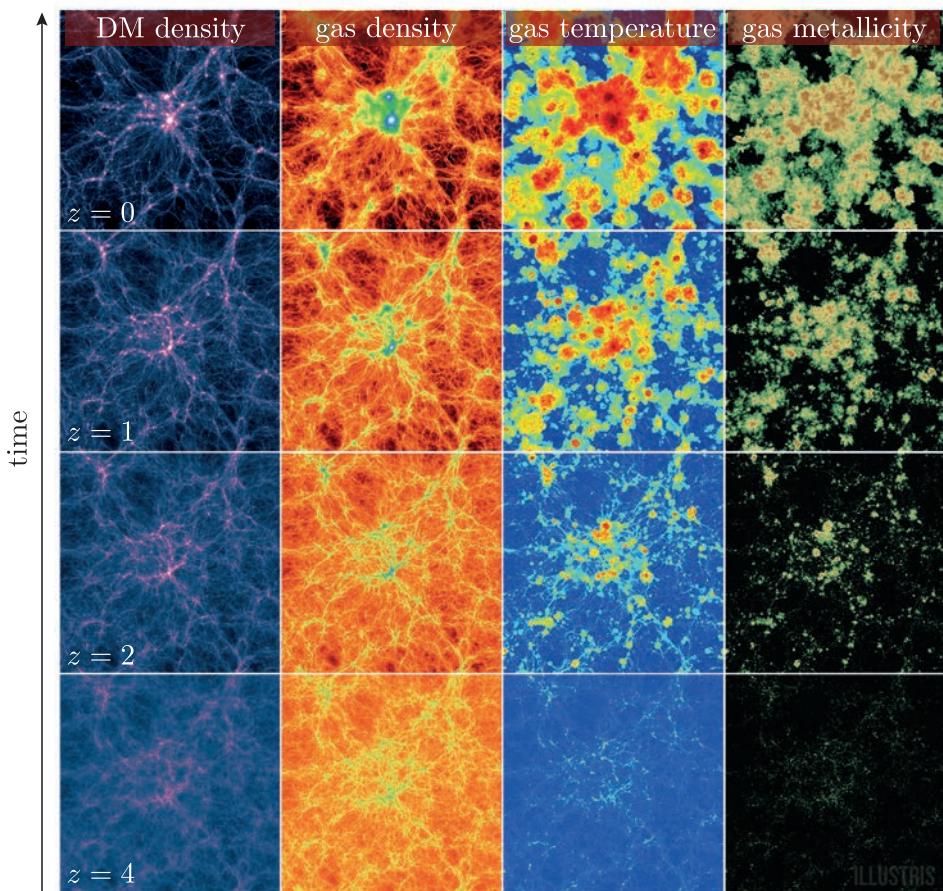


Figure 10.4 The growth of structure in the Illustris cosmological simulation, shown via snapshots of the distribution of four modelled physical properties (dark matter density, gas density, temperature and metallicity) at four different redshifts (z).

In the next chapter you will explore in more detail how baryonic gas evolves within the landscape of the dark-matter structures whose evolution we have examined in this chapter. By incorporating additional physical processes, simulations like the one shown in Figure 10.4 can trace when, where and how individual galaxies form within the cosmic web of dark matter and gas.

10.3.3 Testing simulation predictions

The results of cosmological simulations such as the one shown in Figure 10.4 are constantly being compared with a wide variety of observations of galaxies in the real Universe. These comparisons are extremely successful in many cases, but can also reveal interesting disagreements that are used to identify areas where our theories are incomplete or perhaps entirely wrong.

Modern simulations are thought to model the formation of dark-matter structure extremely well. The dark-matter structure in simulations, and in the real Universe, can be characterised by the **halo mass function**, which is the distribution of dark-matter halos according to mass. Figure 10.5 shows the number of halos, $N (> M)$, in a fixed volume of a large cosmological simulation.

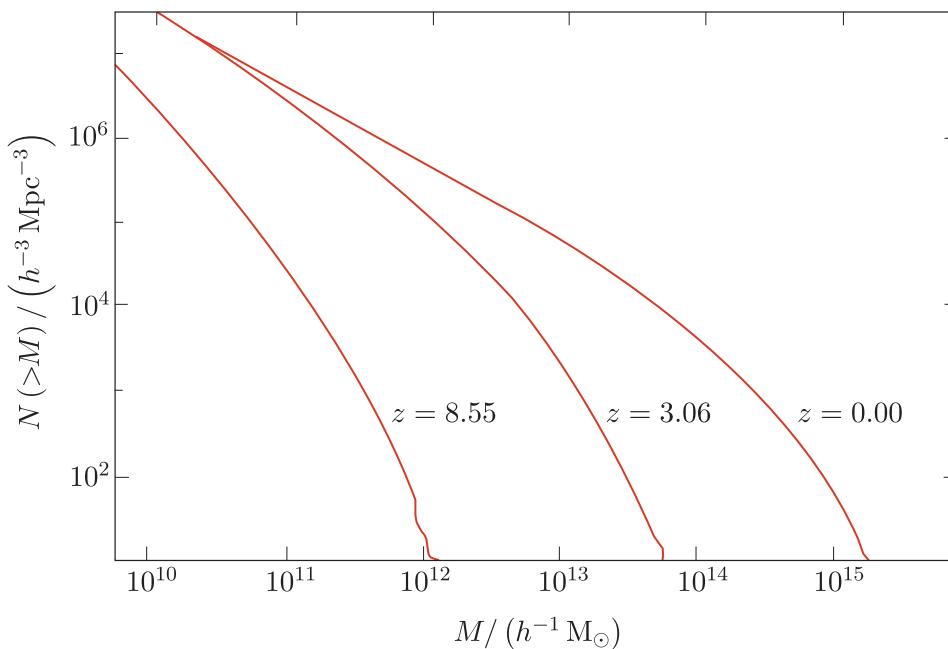


Figure 10.5 The halo mass function at three redshifts from the Millennium Simulation, showing the number of dark-matter halos (N) above a given halo mass (M).

- What can you conclude from Figure 10.5 about the relative numbers of low- and high-mass halos and how these change with time?
- At all three redshifts there are many more low-mass halos than high-mass ones. As the age of the Universe increases (redshift decreases), halos become increasingly massive. Halos of the size of galaxy clusters (10^{14} to $10^{15} M_\odot$) only form at late times (after $z = 3$).

The halo mass functions from simulations can be directly compared with observations. Figure 10.6 shows a comparison of the galaxy cluster mass function measured from X-ray observations with a model obtained by fitting a function to cosmological simulation results similar to those in Figure 10.5.

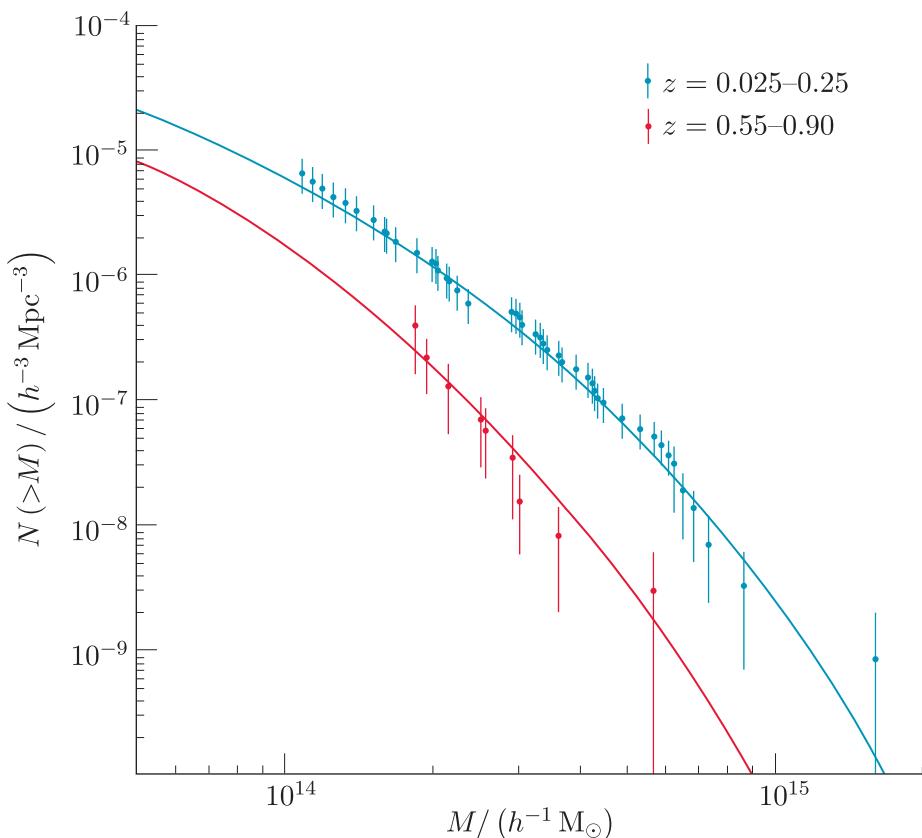


Figure 10.6 A comparison of the observed cluster mass distribution at two redshifts (blue and red data points), with solid lines indicating model fits based on simulation results for the same redshifts.

It is an impressive success both of modern cosmology and developments in numerical simulation methods that it is possible to predict the distribution of halo masses in the present-day Universe on the basis of evolution under gravity of tiny density perturbations created at very early times. The next chapter will take the story further by exploring how galaxies form within the structures we have explored here.

10.4 Summary of Chapter 10

- The expansion of space (Hubble flow) causes galaxies to recede from each other, except on scales of a few megaparsecs, where gravity is strong enough to counteract this expansion.
- The **Local Group** is our gravitationally bound set of galaxies, comprising the galaxies within about 1.3 Mpc of us. It exists as part of larger structures that have formed from the anisotropies present before the formation of the CMB.

- The simplest model for gravitational collapse of structure is based on consideration of the **Jeans criterion**. In this model, objects with masses greater than the Jeans mass can collapse:

$$M_J = \frac{9}{4} \times \sqrt{\frac{1}{2\pi n}} \times \frac{1}{\langle m \rangle^2} \times \left(\frac{k_B T}{G} \right)^{3/2} \quad (\text{Eqn 10.1})$$

- Overdense regions (halos) will start to collapse on the **free-fall timescale**

$$t_{\text{ff}} \approx 18.45 \text{ hours} \left(\frac{\rho}{1 \text{ kg m}^{-3}} \right)^{-1/2} \quad (\text{Eqn 10.7})$$

but the collapse timescale will be increased by the expansion of the Universe, which acts to slow collapse.

- Accounting for both gravitational collapse and expansion of the Universe, the growth of an overdensity, $\delta = (\rho - \langle \rho \rangle)/\langle \rho \rangle$ (Equation 10.3), can be modelled to obtain the following equation:

$$\ddot{\delta} + 2H\dot{\delta} = 4\pi G\langle \rho \rangle \delta \quad (\text{Eqn 10.15})$$

- Overdensities initially grow at a rate of

$$\delta(t) \propto t^{2/3} \propto a(t) \propto \frac{1}{1+z} \quad (\text{Eqn 10.16})$$

- Overdensities in the CMB, originating from the oscillating photon–baryon fluid, aren't large enough to grow the observed galaxy structure. This motivates the need for (non-baryonic) dark matter, of a form that is not coupled to radiation, to drive the collapse of structure.
- Dark-matter halos begin to grow at matter–radiation equality. Baryonic gas can start to collapse into the halos after recombination, when radiation pressure no longer opposes its collapse.
- A collapsing gravitational halo will rebound and virialise to form a semi-stable **bound halo**.
- The most realistic model considers collapse in three dimensions. Collapse first progresses as sheets (i.e. **Zeldovich pancakes**), then **filaments**, and then halos. This creates a foam-like structure, consistent with the observed cosmic web.
- Observed structure in the Universe can be recreated using computer simulations. These simulations take an initial model universe and use physical laws to move that universe forward in time.
- The simulations use different computational techniques to keep track of parcels of material. These include ***N*-body simulations**, hydrodynamical **grid simulations** (fixed or **moving-mesh**, with or without **nested sampling**) and **smoothed-particle hydrodynamic (SPH) simulations**.
- Simulations predict many observable quantities that enable simulation results to be tested. The **halo mass function** is one important, testable prediction of cosmological simulations.

Chapter 11 Formation of stars and galaxies

In Chapter 10 we examined the collapse of dark-matter overdensities, which form halos over a very wide mass range ($\sim 10^5\text{--}10^{16} M_\odot$), explored simple models for dark-matter overdensity collapse and compared sophisticated numerical simulations with observations. We briefly considered the behaviour of ordinary (baryonic) gas in the early Universe and concluded that the gravitational collapse of gas must have taken place after dark-matter overdensities were able to grow substantially.

In this chapter we will explore what happens to the baryonic matter in more detail. We will investigate the processes that follow the collapse of dark-matter halos and how these processes can explain the present-day distribution of stars and galaxies in the local Universe. The aim of this chapter is to provide a relatively high-level overview of how the observable structure in the present-day Universe is understood to have evolved. We note that the physics of star formation is a complex topic in its own right and a full treatment is beyond the scope of this module.

Objectives

Working through this chapter will enable you to:

- explain how baryonic material evolves in the early Universe, and the differences and similarities in the evolution of baryonic and dark matter
- describe the cooling processes in the Universe and their implications for how stars and galaxies formed
- summarise how matter is cycled between stars and the interstellar medium
- discuss the main processes affecting galaxy evolution, including mergers, stellar and black-hole feedback and chemical changes
- summarise how large catalogues of galaxies are used to test models of structure formation and galaxy evolution, and how galaxy populations are used as cosmological probes.

Online resources: properties of galaxies

This chapter assumes a basic understanding of galaxies, including the structure of galaxies like the Milky Way, the main galaxy types, and a basic understanding of the lives of stars. If you have not previously studied the properties of galaxies, you may find the online module resources on this topic helpful, which are taken from our Stage 2 curriculum.

11.1 From baryonic gas to stars

In previous chapters we have seen how primordial fluctuations in the early Universe imprinted structure in the CMB radiation, and led to the growth of a cosmic web of dark-matter halos. In this section we will examine how the baryonic gas collapses into the dark-matter halos, and then cools and compresses to reach densities capable of forming the first stars.

11.1.1 Baryons in dark-matter halos

We have also seen that the mass range of dark-matter halos has some correspondence with the mass range we observe for galaxies in the present-day Universe. This suggests that the collapse of ordinary matter (baryonic gas) proceeds in a similar way to that of the dark matter.

- Explain why the baryonic gas will not collapse in *exactly* the same way as the dark matter.
- The baryonic gas will be attracted to the growing dark-matter overdensities through gravity. However, as we saw in Chapter 7, at early times baryons are coupled to radiation to form a photon–baryon fluid, which oscillates in a landscape of dark-matter overdensities and underdensities. This prevents the baryons from immediately collapsing in the same way as the dark matter.

Figure 11.1 shows how overdensities of dark matter and baryons evolve with time. (Recall from Section 10.1.2 that δ is defined as the local density relative to the background mean density.)

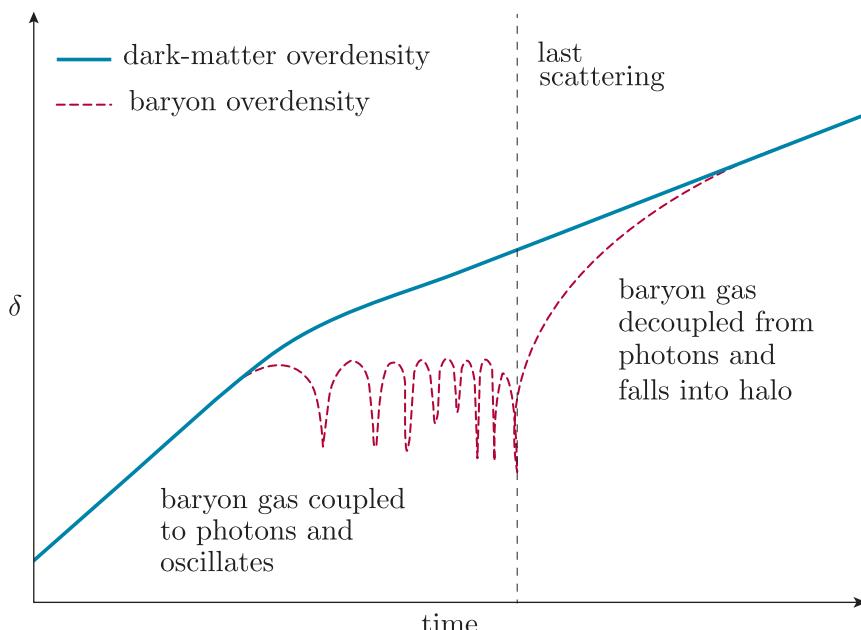


Figure 11.1 A schematic illustration of the evolution of an overdensity of dark matter and an overdensity of baryonic gas at early times.

The dark-matter overdensity grows continuously with time as gravity pulls more dark matter together. In contrast, after a short period of initial growth, the baryonic overdensity starts to oscillate as part of the baryon–photon fluid due to the opposing effects of radiation pressure and gravity. However, the decoupling of photons from baryons at the epoch of last scattering (the vertical dashed line in Figure 11.1) means that the supporting pressure of radiation is removed from the baryonic gas. This enables the baryons to stop oscillating and collapse further into the dark-matter overdensities. The baryonic overdensities are now able to grow much more rapidly, and they quickly evolve to track the dark-matter overdensities more closely.

The baryonic gas within dark-matter halos is the material from which stars and galaxies will form. However, although the gravitational attraction of the dark matter is strong, star formation does not take place immediately. Structures continued to grow in the early Universe through the period sometimes called the dark ages, during which no stars were present. This was a comparatively long period in the Universe’s early history, beginning at the epoch of last scattering ($z \approx 1100$, when the Universe was around 380 000 years old) and lasting for about 200 million years (until $z \sim 15–20$).

Figure 11.2 shows part of a simulation of the density structure in dark matter and in gas for a short period following the end of the dark ages. The bright regions indicate the highest densities. Similar structures are visible in the two panels, demonstrating that the baryons and dark matter trace the same structures across the redshift range shown.

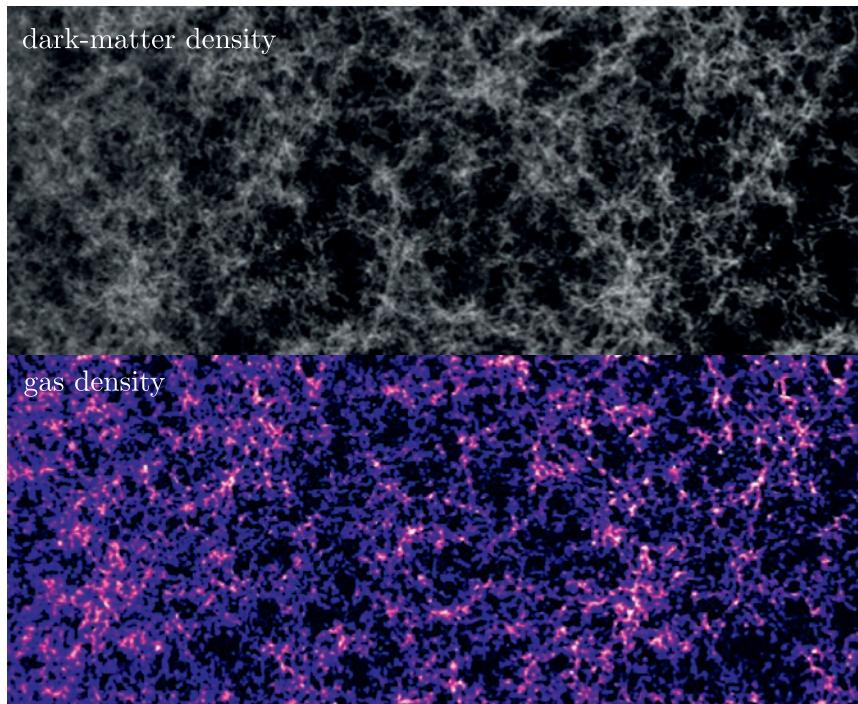


Figure 11.2 The structure of dark matter (top) and baryonic gas (bottom) in the early Universe towards the end of the dark ages, with redshift decreasing from $z = 16$ at the left to $z = 5$ at the right.

11.1.2 Forming structure on galaxy scales

To understand why baryonic collapse does not immediately proceed to turn all of the gas into stars, recall the virial theorem (Equation 9.14).

A virialised system is one where the total average kinetic energy and total average gravitational potential energy are balanced such that

$\langle E_k \rangle = \frac{1}{2} \langle E_p \rangle$, but where energy is constantly exchanged between the two as particles move towards or away from the centre of mass.

The constant exchange of kinetic and gravitational potential energy can be understood in terms of kinetic theory. Clumps of matter fall into the gravitational potential well and collide, causing the macroscopic kinetic energy of the cloud to be distributed among its individual particles, increasing their random motions – what we would measure as an increased temperature. As gravitational collapse proceeds, smaller clumps formed within the baryonic gas will move together with high bulk velocities and cause shock waves and thus heating when they collide. Section 9.3.1 explained that the gas temperatures in these clumps can be as high as 10^8 K. It is the pressure from this hot gas that prevents localised collapse of baryons into galaxies and stars. The gas needs to cool so that galaxy-scale and stellar-mass clumps are below the Jeans mass (Section 10.1.2).

- What mechanism can act to reduce the temperature of baryonic matter to allow stars and galaxies to form?
- Baryonic matter differs from dark matter in that it interacts with itself and with photons, which means that it can radiate heat. Therefore, if the environment is transparent, then radiation can carry away energy from the collapsing cloud. This cooling process reduces the gas pressure and allows the cloud to reach densities where stars can form.

11.1.3 Gas cooling

Section 10.2.1 stated that dark-matter halo collapse can occur on gigayear timescales. During the dark ages, from recombination ($z \approx 1100$) to $z \approx 20$ (the exact endpoint is highly uncertain), baryon collapse proceeded slowly and did not reach the densities required to form stars. Although matter and radiation were decoupled at recombination, the CMB radiation continued to bathe matter in heat, which dictated the lowest temperature at which matter could easily exist. However, by $z \approx 17$, the CMB temperature had dropped from 3000 K (when it was emitted) to 50 K, meaning gas in the Universe could potentially cool to these temperatures.

Before considering the potential gas-cooling mechanisms that were present in the early Universe, it is instructive to review the main cooling mechanisms at work in the Universe today. Table 11.1 lists a range of processes that include interactions of electrons and ions, electron transitions within atoms (including **forbidden lines**, which only occur in low-density space plasmas) and vibrations and rotations of molecules. Each process occurs when gas is in a particular, restricted energy range, and each involves the release of energy as radiation in a particular part of the electromagnetic spectrum, as shown in the last column.

Table 11.1 Primary present-day cooling mechanisms for gas in the Milky Way.

Temperature	Main cooling mechanism	Emission band
$>10^7$ K	Free-free (bremsstrahlung)	X-ray
10^7 – 10^8 K	Iron resonance lines	X-ray
10^5 – 10^7 K	Other metal resonance lines	ultraviolet (UV), X-ray
8000 – 10^5 K	Metal forbidden lines (especially C, N, O, Ne)	infrared (IR), optical
1000 – 8000 K	Atomic transitions (e.g. H, O)	IR, optical
100 – 1000 K	Molecular vibrational and atomic transitions (e.g. O, C ⁺ , H ₂)	far-IR
10 – 100 K	Molecular rotational transitions (especially CO)	sub-millimetre

- How do you think the gas-cooling processes in the early Universe might have differed from the present-day mechanisms given in Table 11.1?
- Baryonic matter in the early Universe was comprised essentially of hydrogen and helium, without astronomical ‘metals’. This means none of the cooling mechanisms in Table 11.1 that rely on metals could operate, (i.e. metal resonance and forbidden lines, atomic and molecular transitions other than hydrogen and H₂). Hence, gas in the early Universe cooled much more slowly.

Figure 11.3 plots **cooling rate**, Λ_{cool} , as a function of temperature, T , for interstellar gas of different compositions.

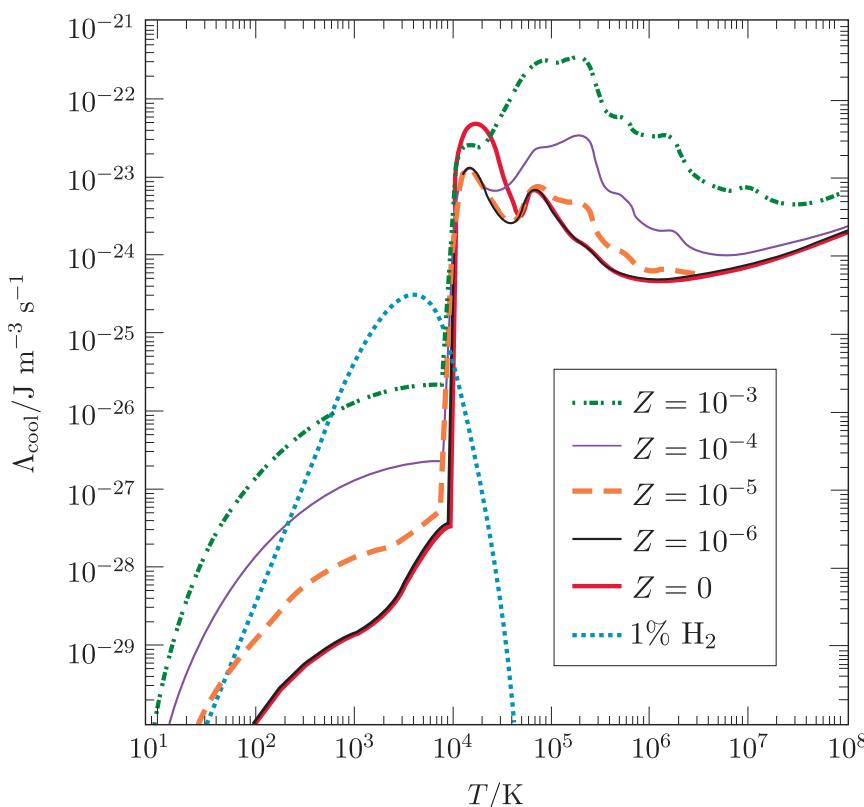


Figure 11.3 Cooling rate, Λ_{cool} , of interstellar gas for different metallicities, Z . Note that the five curves labelled with their metallicity values do not include the effect of cooling by H₂. The cooling rate for interstellar gas containing an H₂ fraction of 1% is shown separately.

- Consider the units of Λ_{cool} in Figure 11.3. What does this suggest about how Λ_{cool} relates to the luminosity of radiation produced?
- The units of Λ_{cool} are $\text{J m}^{-3} \text{s}^{-1}$, equivalent to W m^{-3} . This is the rate at which a cubic metre of material will lose heat energy by radiation, so this defines not only how quickly the material will cool down, but also the luminosity of radiation it will produce.

The lack of effective cooling mechanisms in the early Universe meant the gas pressure remained high. This kept the Jeans mass high also (Section 10.1.2) and meant gas clouds could more easily heat up as they collapsed and dissipated again. An important part of the gas collapse process was the formation of molecular hydrogen, H_2 . In today's Universe, H_2 mostly forms on interstellar dust grains, but in the dust-free primordial Universe it could only happen in a comparatively cool, dense environment where hydrogen atoms would collide frequently at slow enough speeds so that they stuck together. Forming H_2 in this manner released chemical binding energy. This process, **H₂ formation heating**, can heat up gas clouds, helping them dissipate. The ability of gas to cool therefore depends both on its temperature and density.

Figure 11.4 show the relationship between temperature and number density as typical gas clouds cool, for two different chemical compositions: the primordial composition of zero metallicity Z defined in Equation 9.11 and solar composition. A gravitationally contracting cloud will increase in density, moving from left to right on the diagram, following a track that is determined by the interplay of cooling and heating as collapse progresses.

If clumps of gas can cool as they increase in density, then they will fragment into smaller clumps that can go on to form stars. If they cannot cool as they become denser, then they will expand and not go on to form stars. When the gas becomes opaque at the highest densities, radiation can no longer escape and so gas heats much more rapidly.

Exercise 11.1

Examine the two cooling tracks shown in Figure 11.4.

- (a) Over which number density range on the cooling track for solar composition is the gas cooling, and over which range is it heating?
- (b) What is the Jeans mass corresponding to the conditions where stars of solar composition are most likely to form? How does this compare to typical present-day stars?
- (c) Using similar considerations, what would you expect to be the typical mass of the first stars?

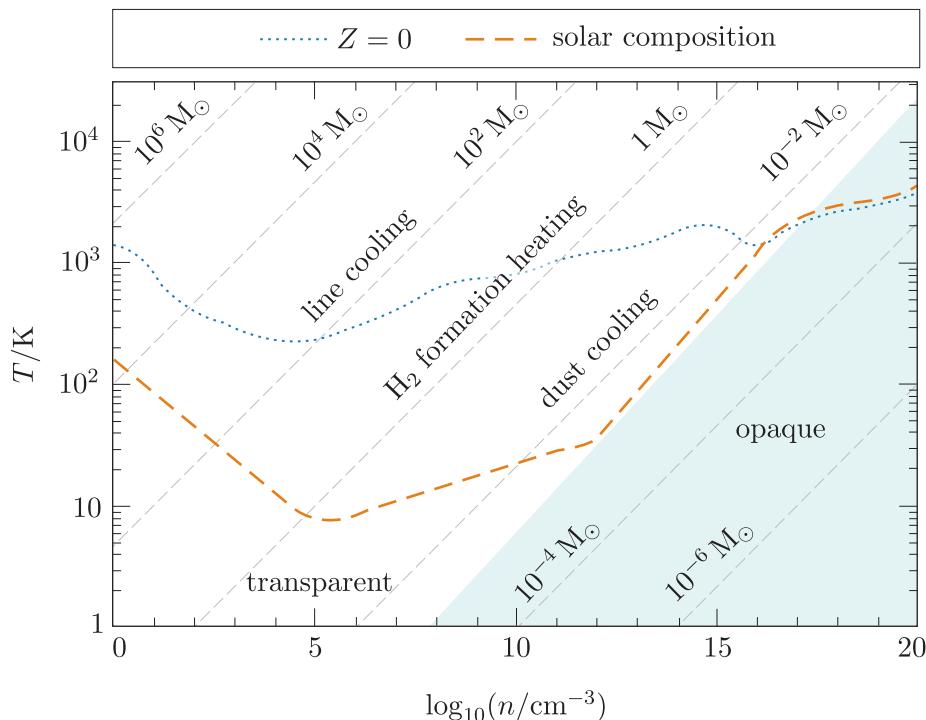


Figure 11.4 Cooling tracks of interstellar gas of primordial ($Z = 0$) and solar compositions. Diagonal lines show the Jeans mass (Equation 10.1) of clumps at that number density, n , and temperature, T . The main cooling and heating mechanisms at each mass are indicated. The blue shaded region shows where the gas becomes opaque.

We can work out an approximate timescale on which gas can cool enough so that stars can form. Consider the kinetic energy contained within atomic hydrogen gas, $E_k = \frac{\gamma}{2} k_B T$ per particle, where k_B is the Boltzmann constant, T is temperature and γ is the number of degrees of freedom. ($\gamma = 3$ for atomic gas and $\gamma = 6$ for molecular hydrogen, H_2 . Note that γ here is unrelated to the Lorentz factor, for which the same symbol is used.) The total kinetic energy per unit volume is nE_k where n is the number density of particles. Then, assuming a constant cooling rate, the **cooling timescale** (which can be considered equivalent to the minimum timescale on which stars could form) is given as:

Cooling timescale

$$t_{\text{cool}} = \frac{nE_k}{\Lambda_{\text{cool}}} = \frac{\gamma nk_B T}{2\Lambda_{\text{cool}}} \quad (11.1)$$

The following exercise explores further how the concept of cooling timescales can be used to compare star formation today to formation of the first stars.

Exercise 11.2

At recombination ($\approx 380\,000$ years after the big bang), the baryonic number density in the Universe was $n \approx 3 \times 10^8 \text{ m}^{-3}$ and the temperature was $T \approx 3000 \text{ K}$. Use this information together with Figure 11.3 and Equation 11.1 to answer the following questions about how different assumptions affect the cooling timescales.

- First make the assumption that the gas at recombination has a metallicity of zero ($Z = 0$) and contains no molecular hydrogen. Estimate the time taken for stars to form, and therefore the age of the Universe at which the first stars could appear.
- Now assume instead that the metal-free gas does contain molecular hydrogen at the level of 1% H_2 and carry out a similar calculation to estimate a timescale on which stars form.
- Comment on what the results of parts (a) and (b) suggest about the role of H_2 in how gas cools to form the first stars.
- By considering the time-evolution of both the Universe and a collapsing cloud of baryons within it, state some of the reasons why the ages you have calculated are estimates.

Exercises 11.1 and 11.2 suggest that the first stars took millions of years to form (detailed calculations give the 200 million years mentioned earlier in the chapter); the exercises also imply that the first stars were extremely massive. The more massive a star, the shorter it lives: if the first stars were very massive, they would have burned brightly but would have exploded very quickly (in less than 10^6 years) as supernovae. We wouldn't see any of these stars at low or medium redshifts today. Searches for these first stars must therefore look to the highest-redshift galaxies in the Universe: we cannot observe individual stars at such colossal distances, but must rely on indirect tracers. The study of these first stars is therefore a challenging and active topic of research. We call these still-hypothetical massive first stars **Population III** stars to differentiate them from the two populations of stars we already observe in the nearby Universe: older **Population II** and younger **Population I** stars.

11.1.4 Reionisation and Strömgren spheres

Massive (blue) stars generate a lot of UV radiation at wavelengths shorter than the **Lyman limit** (91.2 nm). Photons at wavelengths below the Lyman limit can ionise hydrogen atoms directly from their ground state, so the gas in the interstellar medium (ISM) around the first stars became ionised. This was a crucial point in the history of the Universe, known as **reionisation**, and represented the point at which the complex physics of baryonic matter became very important for the observable behaviour of the Universe.

The ionised bubble of plasma surrounding an individual star is known as a **Strömgren sphere**, and will grow outwards from the star. We can model the growth of a Strömgren sphere by considering the rate at which ionising photons are produced by a source, Q , and the number density of material the source is ionising, n . For simplicity, we will assume a cosmos of pure neutral hydrogen ($n = n_{\text{H}}$).

In a relatively diffuse medium, most photons should reach the edge of the ionised region and cause the ionisation of a new atom; the radius R of the Strömgren sphere will expand at a rate of

$$\frac{dR}{dt} \approx \frac{Q}{4\pi R^2 n_{\text{H}}} \quad (11.2)$$

However, in dense regions, some of the ionised atoms will recombine with free electrons. These recombined atoms will then need to be ionised again to maintain the Strömgren sphere, or it will collapse back to neutrality. Mathematically, this effectively decreases Q .

The recombination rate depends on the probability of ion–electron interactions, which is proportional to the ion and electron densities (n_{i} and n_{e} , respectively) and the probability of the two recombining when they meet – the **recombination coefficient**, $\alpha(T)$. Note that $\alpha(T)$ is temperature-dependent, since it is more difficult for ions to capture faster electrons. Note also that at this later cosmological epoch the term ‘recombination’ is more apt than the use of the same term to describe the formation of the first atoms: during reionisation the atoms are indeed recombining having previously been ionised.

The average time it takes an ion to capture an electron, t_{r} , is given by:

$$t_{\text{r}} = [n_{\text{e}}\alpha(T)]^{-1} \quad (11.3)$$

Including the recombination coefficient in Equation 11.2 gives:

$$\begin{aligned} \frac{dR}{dt} &\approx \left[Q - \frac{4}{3}\pi R^3 n_{\text{i}} n_{\text{e}} \alpha(T) \right] \frac{1}{4\pi R^2 n_{\text{H}}} \\ &\approx \frac{Q}{4\pi R^2 n_{\text{H}}} - \frac{n_{\text{i}} n_{\text{e}} \alpha(T) R}{3n_{\text{H}}} \end{aligned} \quad (11.4)$$

By setting $dR/dt = 0$ and rearranging Equation 11.4, we can determine the radius R_{max} where recombination balances out ionisation. Hence, the final (i.e. maximum) radius of the Strömgren sphere is

$$R_{\text{max}} \approx \left[\frac{3Q}{4\pi n_{\text{i}} n_{\text{e}} \alpha(T)} \right]^{1/3} \quad (11.5)$$

In general, as the sphere expands, recombination increases and the growth slows down. By integrating Equation 11.4 we see the radius changes as:

Radius of a Strömgren sphere

$$R(t) \approx R_{\text{max}} [1 - \exp(-t/t_{\text{r}})]^{1/3} \quad (11.6)$$

Exercise 11.3

Consider the following scenarios:

- (i) a very massive, young star embedded in its host galaxy's ISM where $Q = 10^{50} \text{ s}^{-1}$ and $n_{\text{H}} = n_{\text{i}} = n_{\text{e}} = 10 \text{ cm}^{-3}$
- (ii) a galaxy of 10^5 such stars surrounded by a neutral medium at $z \approx 17$ where $Q = 10^{55} \text{ s}^{-1}$ and $n_{\text{H}} = n_{\text{i}} = n_{\text{e}} = 10^{-5} \text{ cm}^{-3}$.

Assuming $\alpha(T) \approx 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1}$ in each case, for both scenarios calculate:

- (a) the maximum radius of the Strömgren sphere
- (b) the timescale for atoms recombining
- (c) the size of the Strömgren sphere after 10^6 years.

Exercise 11.3 shows that even the largest, brightest stars can only ionise a small part of their galaxy's ISM. In contrast, a galaxy of large, bright stars can completely ionise the **intergalactic medium** (IGM) over a much larger region than the size of the galaxy itself. The primary reason is that recombination timescales inside galaxies are short (much shorter than the timescales of stellar evolution), whereas in the IGM between galaxies, an ionised atom is unlikely to ever meet an electron and become neutral again. Consequently, the galactic ISM will remain mostly neutral, with pockets of ionisation generated by localised star formation. Meanwhile, the ionisation front from the first galaxies may have travelled through intergalactic space at close to the speed of light, and the IGM remains almost completely ionised to this day.

The calculations of recombination timescales and the evolution of Strömgren spheres outlined in this section are approximations for several reasons. These include:

- although most of the matter in the Universe is hydrogen, other elements contribute too
- recombination releases photons, some of which will contribute to maintaining ionisation
- if ionisation occurs sufficiently quickly, then lower-energy photons can significantly contribute to it
- the hot, ionised gas will be over-pressured compared to the surrounding neutral gas, so will expand of its own accord
- the edge of the Strömgren sphere can become poorly defined, because the mean free path of particles is long
- Q and T are liable to change as stars evolve
- the expansion of a Strömgren sphere is fundamentally limited by the speed of light.

The period when the Universe was filled with intense radiation and returned to a largely ionised environment is referred to as the **epoch of reionisation**. It lasted from (very approximately) $z \sim 20$ to $z \sim 6$, and transformed the baryonic Universe from a cold, neutral gas (containing density fluctuations) into today's Universe, with dense, neutral interstellar media within galaxies, but hot, ionised intergalactic media outside of galaxies. Studying the epoch of reionisation and determining the redshifts at which it occurred place important constraints on the evolution of the Universe: this is a major goal of telescopes such as *JWST*.

Online resources: reionisation simulations

The snapshots in Figure 11.5 were taken from a video rendering of the full numerical simulation. A link to the simulation can be found in the online module resources.

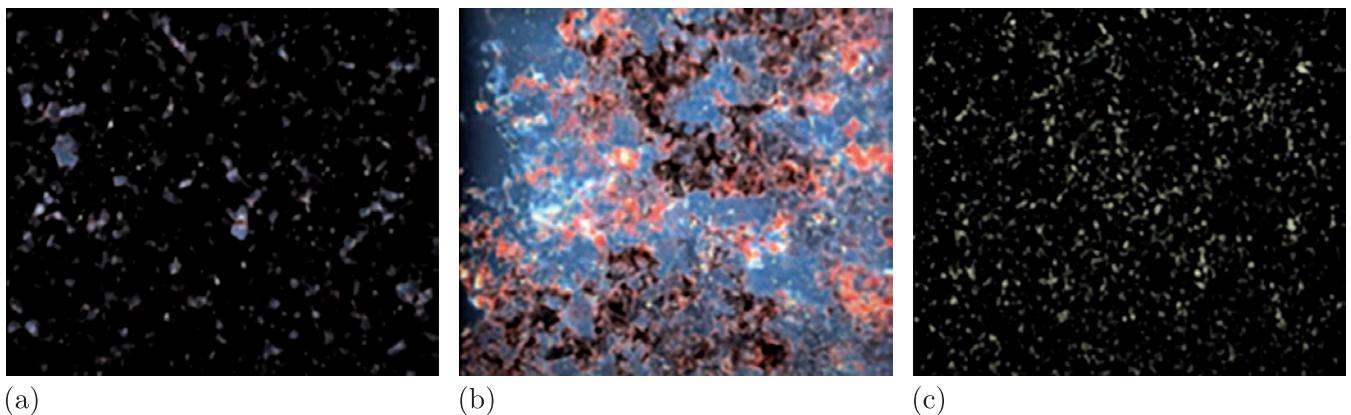


Figure 11.5 Snapshots from a simulation of reionisation, showing its approximate appearance in UV light at three different redshifts from (a) $z \sim 15$ to (b) $z \sim 10$ to (c) $z \sim 6$ (note that the exact redshifts it reaches each stage are not well determined). The Universe is initially opaque to UV with isolated Strömgren spheres. These spheres then merge together and the partially ionised gas becomes transparent in some directions. Finally, the gas becomes diffuse and fully ionised: the Universe is then fully transparent to UV, and we can see galaxies throughout the simulation.

11.2 Key processes in galaxy evolution

This section discusses the main factors that affect the structure and chemistry of galaxies. In particular, we will look at the conversion from the hydrogen–helium Universe of the ‘dark ages’ to the bright, galaxy-filled, chemically rich Universe of today.

11.2.1 Galaxy assembly

Figure 11.6 shows the major processes involved in building a galaxy, and how it evolves with time.

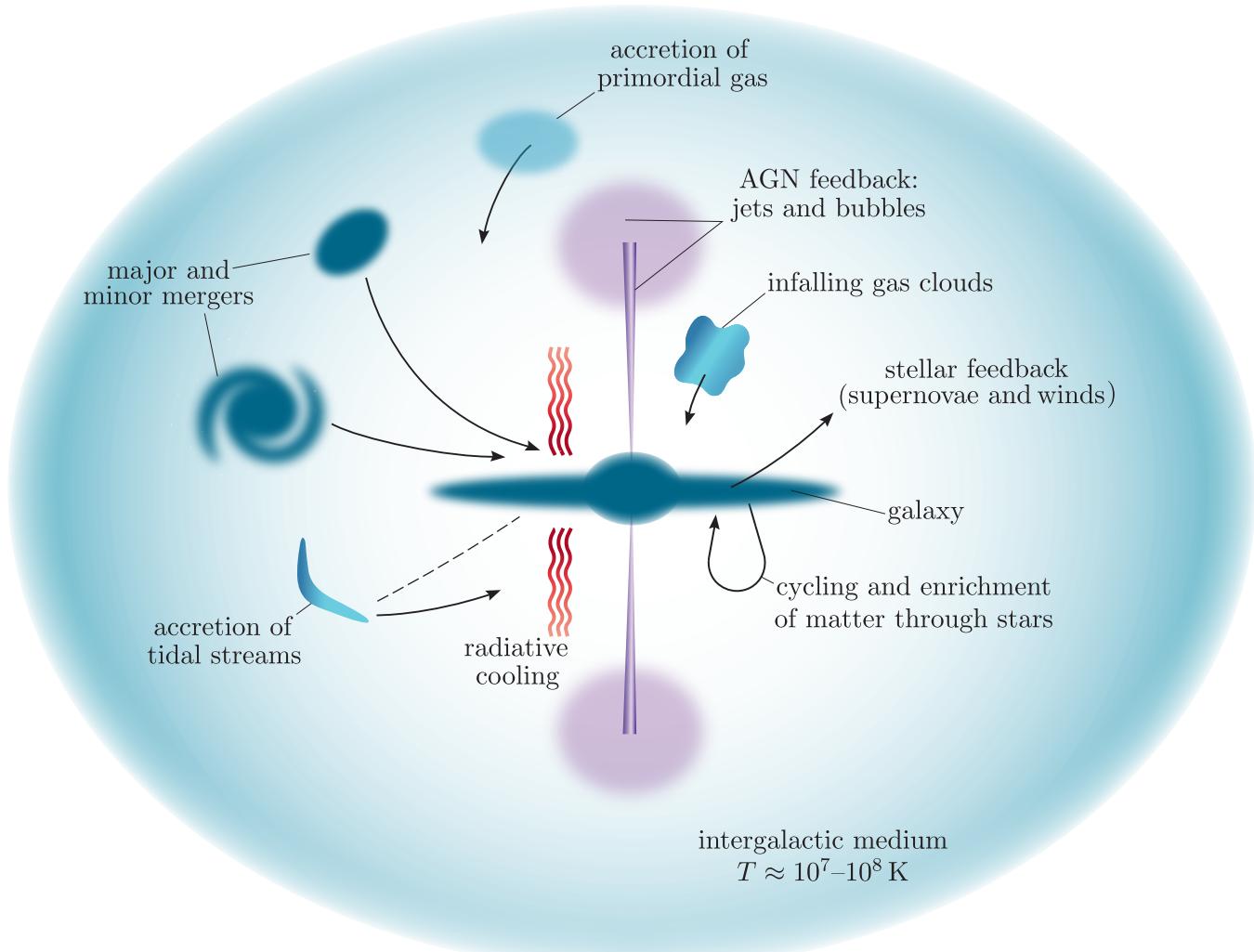


Figure 11.6 The primary methods by which a galaxy interacts with its environment. Note that the infall processes can occur from any direction, but radiation, AGN, stellar and supernova feedback occur primarily in the directions perpendicular to the galaxy disc.

Galaxy formation begins with the accretion of primordial gas clumps, which can cool enough to form stars. These are the first **protogalaxies**, which evolved to become the galaxies we observe at the highest redshifts. Observations of high-redshift galaxies show these to be mostly irregular in shape and typically much smaller in size than nearer (low-redshift) galaxies that are being observed at a later stage in their evolution.

As galaxies evolve, they continue to fall towards the centres of dark-matter halos, forming groups and clusters of galaxies. Collisions between galaxies can lead to **galaxy mergers** – either a major merger between two galaxies of roughly equal size, or a minor merger if a small galaxy falls into a large one. Galaxies can also be shredded into **tidal streams** if they pass too close to a larger galaxy: a process that we observe in the Milky Way today. Stars from these streams can be accreted onto the larger galaxy indirectly.

As early galaxies evolved, they increased in mass through the accretion of gas, while collapsing via gravity as they cooled. A collapsing galaxy is likely to be irregular in shape, which leads to net rotation because collapse is not completely symmetric. Conservation of angular momentum increases the rate of rotation as the collapse progresses. While gravitational collapse still occurs along the axis of rotation, the effect of rotation is to counteract gravitational collapse in directions perpendicular to the axis. Hence, the matter in rotating galaxies tends to form spirals constrained to a disc.

Major galaxy mergers occur in denser environments, and tend to disrupt the rotation of galaxies and scatter stars, leading to the formation of elliptical galaxies. The most dynamic part of a dark-matter halo is its centre: galaxies here accrete the most gas and undergo the most mergers, so the centre of a galaxy cluster tends to host large elliptical galaxies.

If a galaxy has a massive central black hole and the galaxy centre is rich in gas, it is likely to go through phases of gas accretion onto the central black hole, resulting in an **active galactic nucleus** (AGN). The AGN can also produce jets that push through the surrounding gas and terminate in bubbles called radio lobes. Jets, radio lobes and the processes of stellar feedback annotated on the right-hand side of Figure 11.6 are discussed in the remainder of this section.

11.2.2 The cosmic cycle of matter

Stars play a significant role in shaping the chemical composition of a galaxy. Consider the first stars to form in a galaxy comprised mainly of hydrogen and helium, which are expected to have high masses (Section 11.1.3). The rate at which stars consume fuel increases with stellar mass. Hence, the massive first-generation stars will reach the end of their lives quickly. This allows an appreciable amount of material to be ejected from these stars into the ISM via stellar winds and supernovae remnants.

Crucially, this ejected material is chemically enriched with the products of the stars' nuclear fusion, and so astronomical metals begin to diffuse through the ISM. The next generation of stars will form from this chemically enriched ISM, thus creating a cosmic cycle of matter between the ISM and stars (see Figure 11.7).

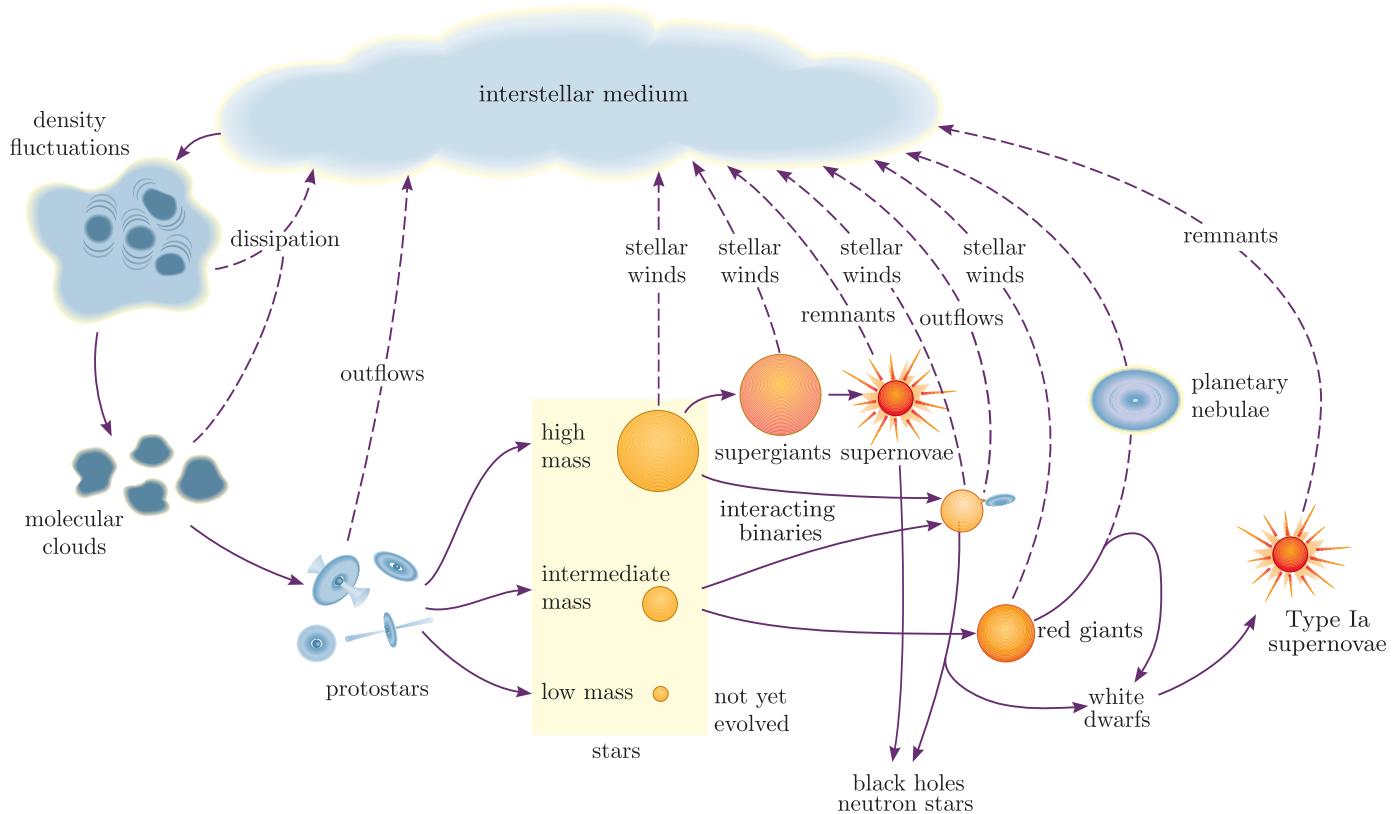


Figure 11.7 The cosmic cycle of matter, showing the main pathways by which matter moves between the ISM and stars and back again. Processes at the left began when stars first formed; processes at the right only began later in the timeline of the Universe.

The cosmic cycle of matter allows us to infer the nature of stars in the early Universe. The oldest stars we see today are Population II stars, formed from gas already enriched by the first generation of Population III stars. The chemical make-up of these early Population II stars reveals the elements ejected by Population III stars, from which we can infer what the Population III stars were like.

The chemical enrichment of the ISM changes with time. As the Universe ages, progressively lower-mass stars start to die. After ~ 35 million years, stars with masses a little less than $8 M_{\odot}$ evolve to become red giants, with increasingly lower-mass stars gradually following them into this stage over time. Towards the end of the lives of the more massive red giants, violent pulsations essentially shake stellar atmospheres into space to form planetary nebulae. During this shedding process, further nuclear reactions take place, involving the capture of neutrons, which enable heavier elements to be synthesised.

Figure 11.8 illustrates, for each element of the Periodic Table, the various – mainly astrophysical – origins for the majority of that element present in the Universe. It shows that both low- and high-mass stars contribute to producing the elements of most importance for everyday life. For example,

carbon and nitrogen are primarily produced in low-mass stars, whereas the most-massive stars are essential to produce a range of other important elements, including oxygen, aluminium, sodium and magnesium.

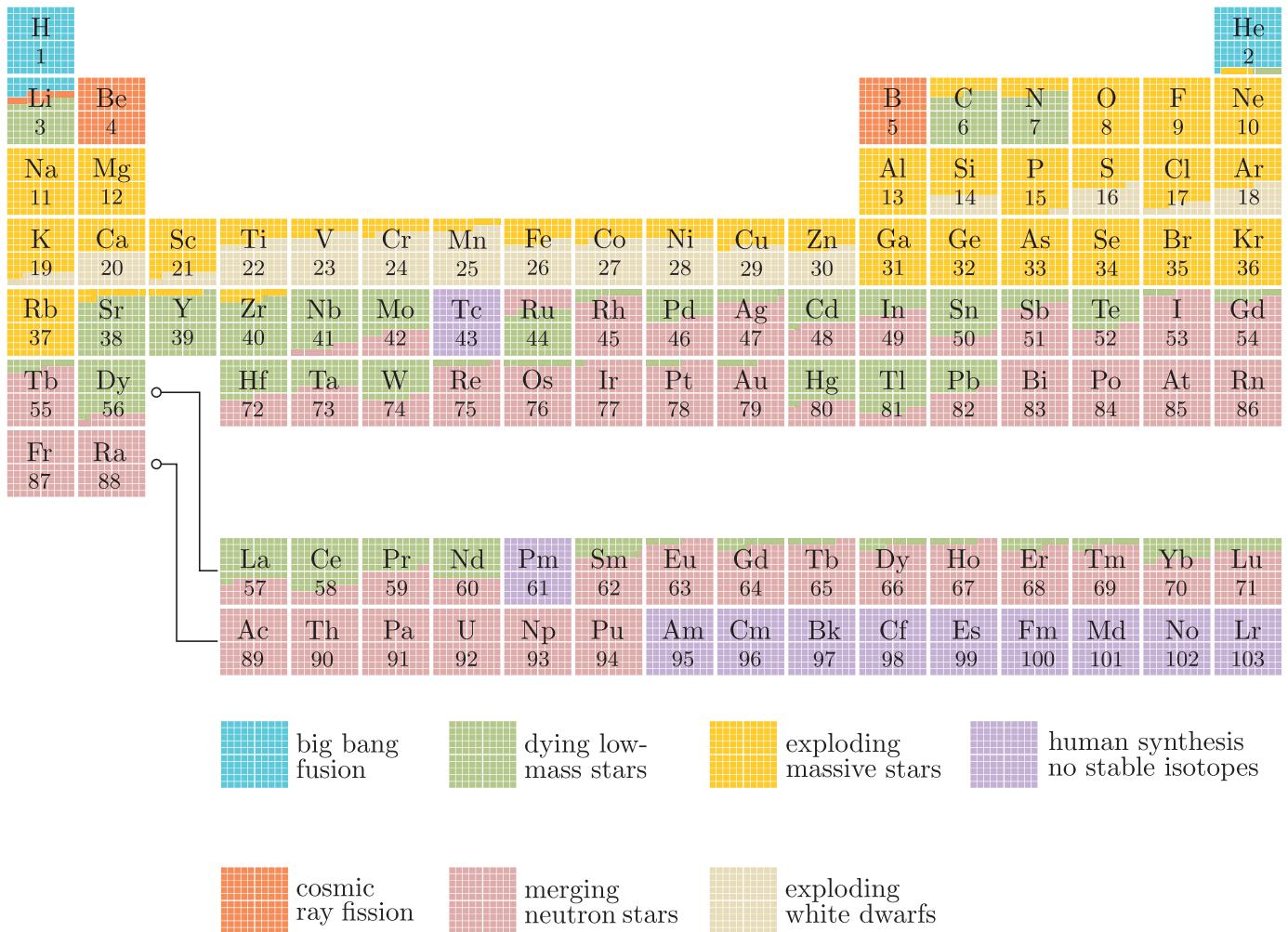


Figure 11.8 The Periodic Table, showing the sources of each element.

As well as enriching the ISM with astronomical metals, stars also provide radiation and kinetic energy to the ISM. This **stellar feedback** (Figure 11.6) can promote or inhibit further star formation. Ejecta from the winds of hot stars, and from supernovae and other outflows, can have enough kinetic energy to escape the gravitational potential of their galaxies. Such ejecta can form bubbles and, especially in disc galaxies, we see these bubbles expanding out of the galaxy as **galactic outflows**, as shown for the galaxy Messier 82 in Figure 11.9. Material from bubbles can subsequently fall back down into the galaxy in different places, replenishing the galaxy's gas reservoir.

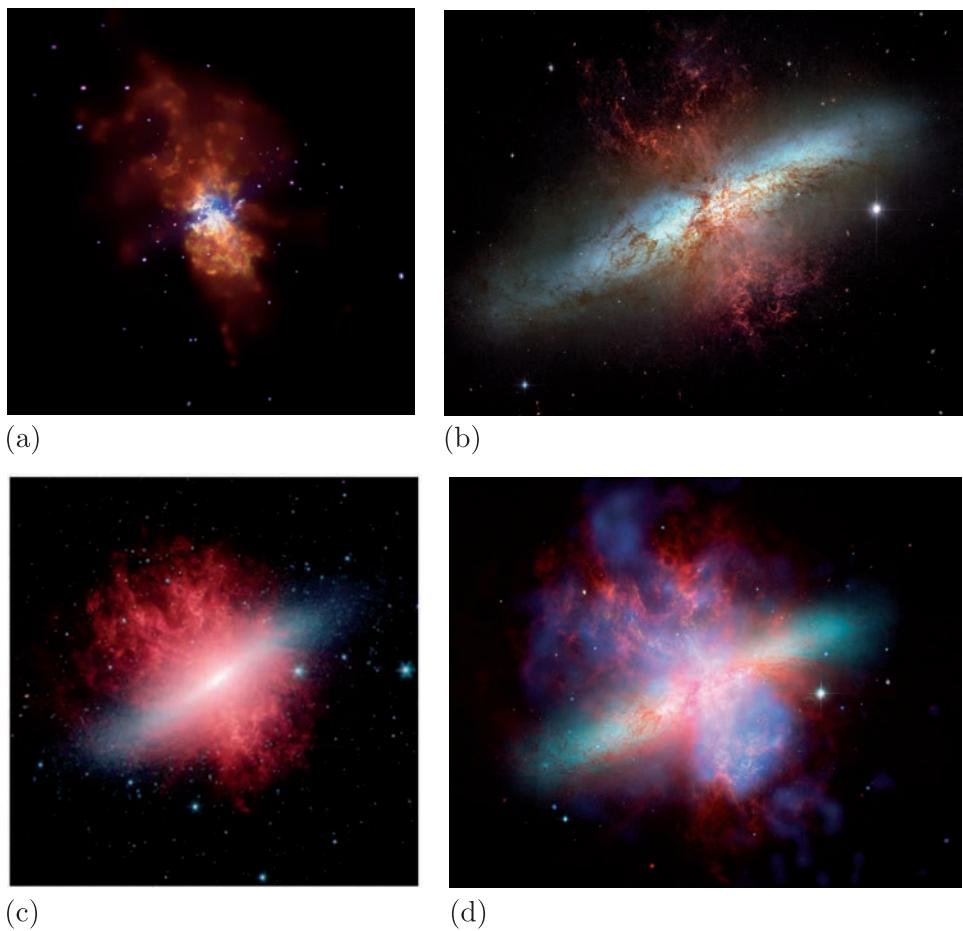


Figure 11.9 Messier 82, as seen in (a) X-ray, (b) optical and (c) infrared, with (d) a colour composite including all three wavebands. The cigar-shaped disc of the galaxy extends from bottom-left to top-right, while the galactic outflow formed by its ongoing episode of rapid star formation is perpendicular to this.

The next exercise involves investigating these feedback processes further.

Exercise 11.4

The escape speed (see Section 3.2.2) for gas at a particular radius R from the centre of a region of mass M is given by

$$v_{\text{esc}} = \sqrt{\frac{2GM}{R}} \quad (11.7)$$

Consider a point $R = 8$ kpc from a galaxy's centre, and assume that the galaxy has a mass of $M = 10^{12} M_{\odot}$ within this radius.

- (a) What is the escape speed from this point?
- (b) How does the escape speed compare with:
 - (i) a 800 km s^{-1} supernova outflow
 - (ii) a 20 km s^{-1} outflow from a dying low-mass star?

- (c) Using Figures 11.6 and/or 11.7, suggest the stellar feedback mechanisms in this galaxy that processes (i) and (ii) contribute to.
- (d) How would your answers to (b) and (c) differ if processes (i) and (ii) were located in a smaller galaxy of $M = 10^8 M_{\odot}$ and $R = 800$ pc. Describe the implications for chemical enrichment in the smaller galaxy.

The information about a galaxy's evolutionary history encoded in the chemical composition of its stars can be used to place limits on how the mass and radius of galaxies have changed over time. This helps astronomers distinguish dark-matter dominated galaxies, which are both massive and compact enough to retain ejecta from supernovae, from smaller systems like globular clusters, which don't contain dark matter and only show small amounts of elements produced in supernovae. The minimum mass of a galaxy, including dark matter, is therefore of order of the most-massive globular cluster: $\sim 10^7 M_{\odot}$.

11.2.3 Feedback from stellar remnants and supermassive black holes

We have seen from the cosmic cycle of matter that stellar winds and supernovae can return a lot of mass from stars to the ISM, but a fraction of matter will remain locked in compact **stellar remnants** (namely black holes, neutron stars or white dwarfs). Since not all mass is recycled, the amount of gas in a galaxy will decline over time. Therefore, unless the gas-depleted galaxy can accrete more gas by merging with a gas-rich galaxy, the galaxy will eventually stop forming any young, hot, blue stars entirely and become 'red and dead'. When we examine high-redshift galaxies, we find that star formation peaked around 10 billion years ago ($z \approx 2$) and has been gradually declining since.

As galaxies age, the number of remnants increases. These stellar remnants continue to play an important role in shaping galaxies. Over half of stars (especially high-mass stars) are found in binary or higher-order systems, and sometimes two stars can interact. **Interacting binary stars** affect each others' nuclear and chemical evolution. If an evolving star transfers mass onto a compact object, it can trigger outbursts called novae, which add a distinct chemical fingerprint to galaxies.

In dense environments, including the centres of galaxies and stellar clusters, stellar encounters are common. This shares out the kinetic energy of the stars and stellar systems: small stars become kinematically hot (i.e. they have high velocities), while massive stars become kinematically cold (i.e. they have small velocities, but similar amounts of orbital momentum). The relevant mass here is the total mass of the stellar system (i.e. binary stars are slowed like one single star of the same mass). This redistribution of kinetic energy leads to **mass segregation**: massive stars

and binary stars – and any remnants they create – will tend to sink to the centre of their host clusters or galaxies. This encourages massive remnants like black holes to congregate towards the galactic centre.

At the centre of all large galaxies are **supermassive black holes** (SMBHs). The origins of SMBHs are not well known, but they attained masses of 10^6 – $10^8 M_\odot$ within a fraction of a billion years of the big bang. One hypothesis is that they built up via a very rapid version of mass segregation from the remnants of Population III stars. An alternative theory is that supermassive black holes originated in massive ‘seed’ black holes formed by direct collapse of large amounts of matter in the very early Universe.

SMBHs are often surrounded by chaotic environments, including active sites of star formation and orbiting, infalling material. In these active galactic nuclei (AGN), material is actively being accreted by the black hole. The mass energy of this accreted material can power **jets** that channel energy, and energetic particles, outwards from very close to the central black hole, at speeds close to the speed of light. The jets produce **synchrotron radiation**, seen at radio wavelengths, and the large bubbles they create are visible as **radio lobes**, as shown in Figure 11.10. Jets from supermassive black holes transfer substantial amounts of energy into the outer regions of the host galaxies and the wider environment, affecting subsequent star formation.

The energy deposited by an AGN into its host galaxy affects galaxy evolution. This **AGN feedback** can be ‘positive’, promoting star formation by triggering the collapse of gas clouds; or ‘negative’, **quenching** (halting) star formation by either preventing radiative cooling of infalling gas, or ejecting too much gas from the galaxy.

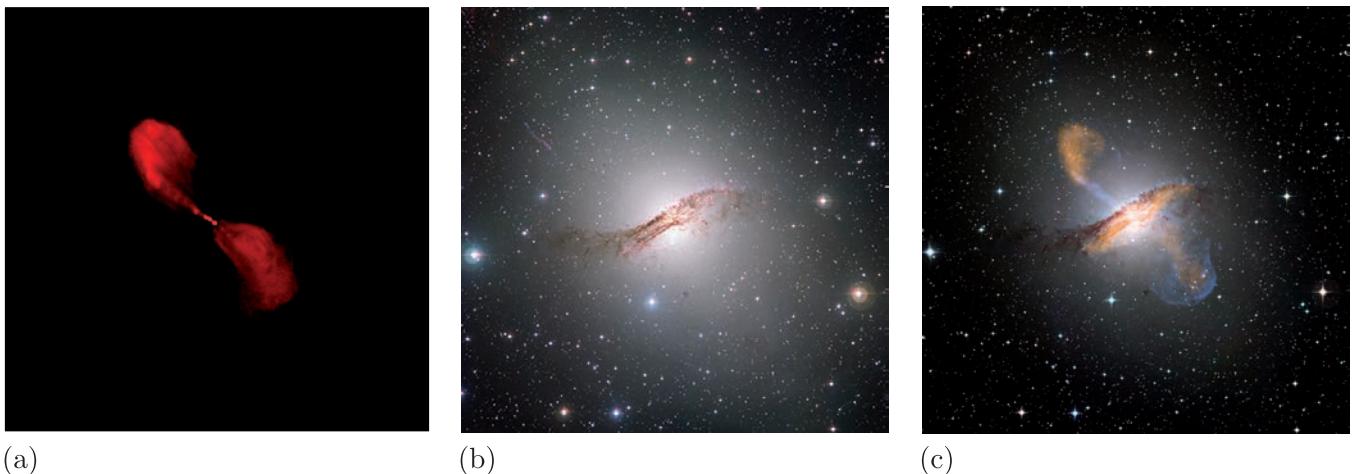


Figure 11.10 (a) Radio jets and lobes of the radio source Centaurus A, which originate in the AGN at the centre of the it hosting galaxy NGC 5128, shown in an optical image in (b), which reveals strong absorption from a dust lane across the centre of the galaxy. Panel (c) shows a super-position of NGC 5128’s optical and radio structure.

11.2.4 Chemical evolution of galaxies

Stellar winds, novae, supernovae, outflows, accretion of infalling gas, and AGN feedback all combine to change the chemical properties of individual galaxies. When galaxies merge, their chemical properties become entwined. However, their individual stellar populations can sometimes remain distinguishable using chemical tracers, such as the ratios of individual elements.

Galactic archaeology describes the process of using chemical tracers to understand the merger histories of our Galaxy and others, and/or to explore how the Universe as a whole became chemically enriched. Chemical tracers can be found either in the spectra of individual stars or those of the interstellar gases in galaxies. Common tracers include oxygen and iron abundances: oxygen is relatively easy to measure in the gas phase, while iron is relatively easy to measure in stars. (Use of these tracers is similar to the methods for investigating primordial elements discussed in Chapter 9.)

Star formation is strongly suppressed when negative AGN feedback is strong, the gas in a galaxy is ejected by outflows, or too much material in a galaxy is locked up in stars or remnants. When star formation is quenched, galaxies will change in appearance from being dominated by hot, young, blue stars, to being dominated by cooler, old, red stars. Hence the colours of galaxies give clues about their evolutionary histories.

In the next exercise you will explore how elemental abundances can be used to investigate galaxy evolutionary histories.

Exercise 11.5

Figure 11.11 (overleaf) shows the relative abundances of oxygen and iron for four galaxies. (Recall the nomenclature $[O/Fe]$ and $[Fe/H]$ described in Section 9.2.1. Values of $[Fe/H] = 0$ or $[O/Fe] = 0$ correspond to solar abundance ratios.)

Use Figure 11.11 together with Figures 11.7 and 11.8 to answer the following questions.

- What are the processes that produce most of the hydrogen, oxygen and iron?
- In what order do these processes occur throughout the Universe?
- Using your answers to parts (a) and (b), explain the common trends in Figure 11.11.
- Describe and explain the differences between the four galaxies shown in Figure 11.11.

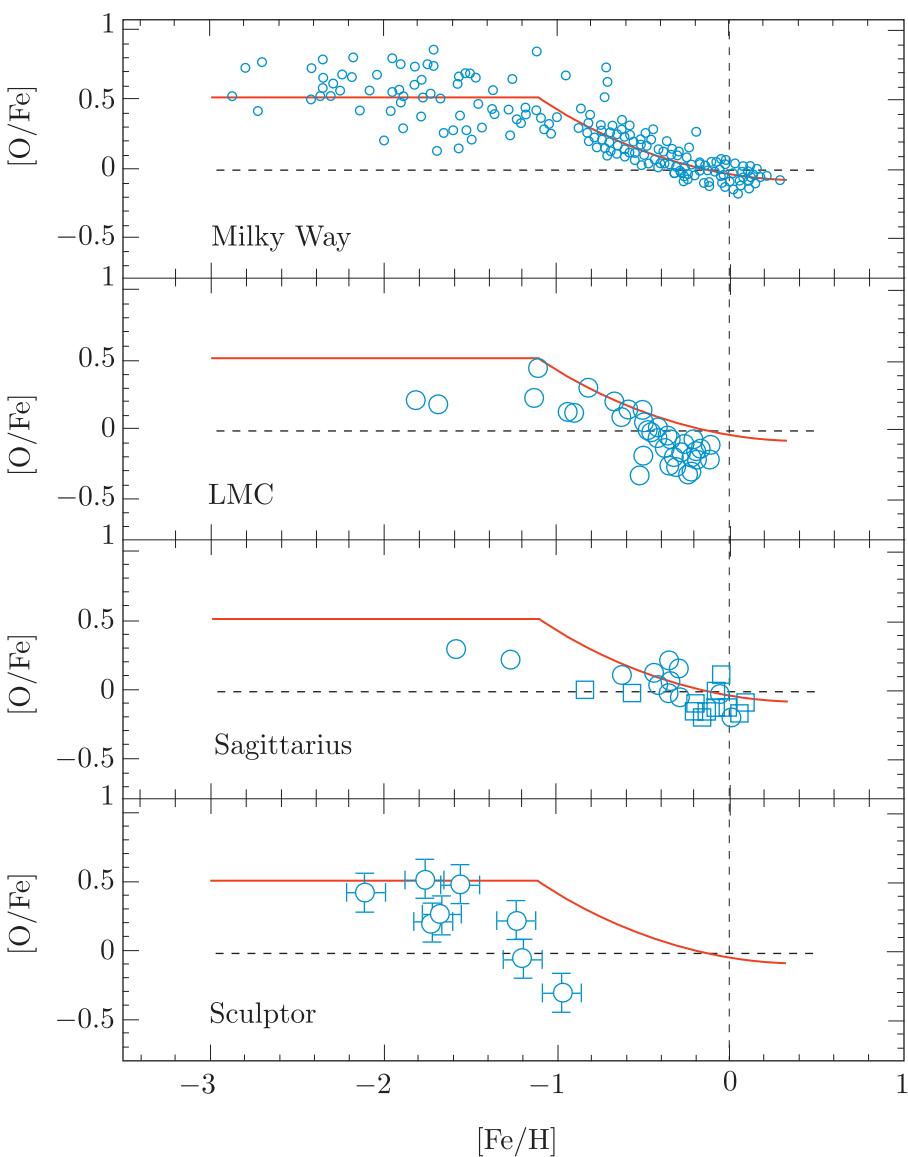


Figure 11.11 Oxygen-to-iron abundance versus iron abundance for stars in the Milky Way and three of its satellite galaxies, the Large Magellanic Cloud (LMC), Sagittarius and Sculptor. The red line shows a model fit to the Milky Way and is identical in all four plots. The vertical dashed line shows the Sun's iron abundance.

11.3 Galaxies as cosmological probes

As we cannot directly observe dark matter, we must rely on baryonic matter in stars, galaxies and galaxy clusters to act as a tracer for all of the mass in the Universe. Knowledge of the processes of galaxy formation and evolution that we discussed in the previous sections is essential to enable reliable conclusions to be drawn about the overall matter distribution from observations of galaxies.

A wide range of observations can be used to trace the matter distribution in the Universe. In earlier chapters you learned about the use of Type Ia supernovae to trace the overall expansion of the Universe, and of galaxy clusters to measure the halo mass function. In this section you will learn about some of the ways that surveys of very large numbers of galaxies can be used to learn about galaxy and structure formation.

11.3.1 Galaxy number counts

Galaxies (and clusters of galaxies) signify localised peaks in the matter distribution of the Universe. The number of these peaks and spacing between them can be used to test whether our understanding of structure formation and galaxy evolution is correct, and to test cosmological models.

In Section 10.3.3 we saw that measurements of the halo mass function could be used to test the predictions of numerical simulations. These types of comparison also allow us to determine which cosmological parameters (e.g. density parameters, neutrino masses, the cosmological constant's equation of state, and many others) are most consistent with the observed properties of the Universe.

The simplest statistical tool that can be used to compare observations with theoretical predictions is that of **number counts**. Here, a histogram is made of the number N of galaxies, or galaxy clusters, in a particular magnitude or flux range. Alternatively, a cumulative histogram can be constructed to show the number of galaxies or clusters brighter than a limiting flux or magnitude.

If the Universe was static and Euclidean (flat), with a relatively uniform distribution of galaxies, we would expect the number of galaxies of a given intrinsic brightness (luminosity) and within a given radius R to be $N($R)$ $\propto R^3$. Since flux is proportional to R^{-2} , this means that the number of galaxies above a flux F should be $N(>F) \propto F^{-3/2}$ or$

$$\log[N(>F)] = -\frac{3}{2} \log(F) + \text{constant} \quad (11.8)$$

Deviations from this expression tell us about the expansion of the Universe and the evolution of structure formation. Since the Universe appears flat, deviations from Equation 11.8 were used as evidence against a steady-state Universe. Today, we use number counts of galaxies to identify how galaxies

have evolved, and number counts of clusters to investigate the formation and evolution of dark-matter halos.

Modern galaxy redshift surveys enable sophisticated tests of cosmological models using measurements of properties such as luminosity and **stellar mass** (total mass contained in stars) for large samples of galaxies across a wide range of redshifts.

11.3.2 Galaxy luminosity and stellar mass functions

The galaxy **luminosity function** and **stellar mass function** are defined similarly to the halo mass function we met in Chapter 10, but they are plotted as a *differential* quantity Φ . If n is the number density of galaxies per cubic megaparsec with luminosities greater than L , then:

Luminosity function

$$\Phi_L = \frac{dn}{d \log_{10} L} \quad (11.9)$$

Similarly, if we instead consider the number density of galaxies above a stellar mass M , then:

Stellar mass function

$$\Phi_M = \frac{dn}{d \log_{10} M} \quad (11.10)$$

Note that $d \log_{10} L$ represents an increment in luminosity on a logarithmic scale, e.g. from 10^9 to $10^{10} L_\odot$. Similarly, $d \log_{10} M$ represents an increment in mass on a logarithmic scale. Hence Φ_L and Φ_M can be thought of as the number of galaxies per cubic megaparsec, per factor of ten in luminosity and mass, respectively.

It is worth noting that, somewhat confusingly, the luminosity function can also be reported in the form of optical magnitudes, which also has the symbol M . Usually it is fairly easy to tell which M is intended as in the case of magnitudes it will have a subscript to indicate a particular telescope filter, whereas a mass function will be reported using units of mass.

Figure 11.12 shows the observed galaxy luminosity function measured in the UV part of the spectrum and the corresponding stellar mass function for a range of redshifts from $z = 0.05$ (shown in dark brown) to $z = 4.0$ (shown in purple).

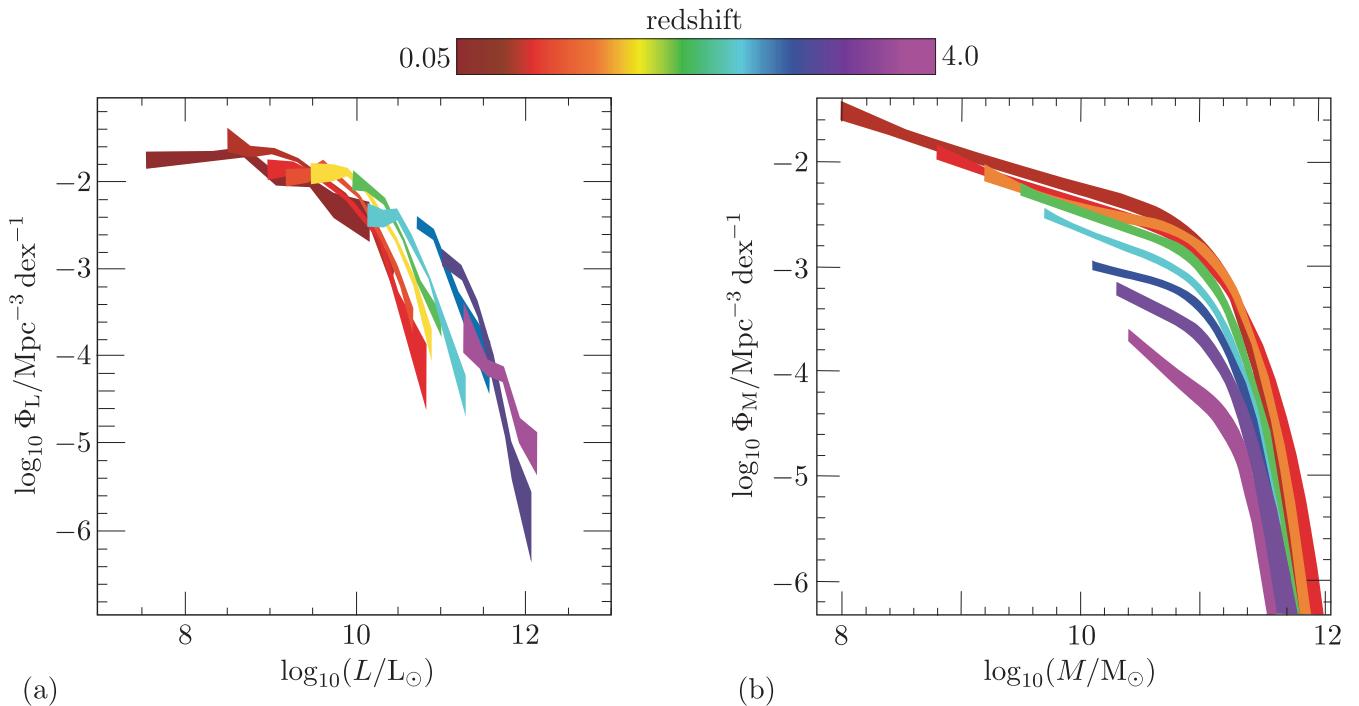


Figure 11.12 (a) The galaxy luminosity function from far-UV observations and (b) the corresponding mass function. In both cases, the vertical axis shows the logarithm of the number of galaxies per cubic megaparsec, per factor of ten (dex) in the horizontal quantity. The coloured bands show the luminosity and mass functions for the indicated redshifts.

- The highest redshift curves drop off at higher luminosity in Figure 11.12a than those at low redshift. Suggest an explanation for this decrease in the number of UV-bright galaxies with time.
- UV-bright stars are hot, blue stars. The far-UV is therefore a star-formation tracer. As the available gas in galaxies has declined, the amount of star formation has dropped from its peak at $z \approx 2$. This causes the point at which the far-UV luminosity function steepens to move left in the plot as we progress towards $z = 0$.
- Similarly, explain the increase in the number of massive galaxies at lower redshifts.
- Galaxy mergers and the continuing inflow of gas increase the mass of the most massive galaxies, meaning the mass function moves upwards with decreasing redshift. (The faint end of the mass function will also move down in time as these galaxies are removed, but faint galaxies are hard to find at very high redshifts.)

In Section 10.3.3 you saw that the observed and theoretical halo mass functions for galaxy clusters are in good agreement. It is also very informative to compare the mass functions for dark-matter halos with those for observed galaxies. Figure 11.13 shows a comparison between the theoretical halo mass function and observations of the stellar mass function, similar to those shown in Figure 11.12b.

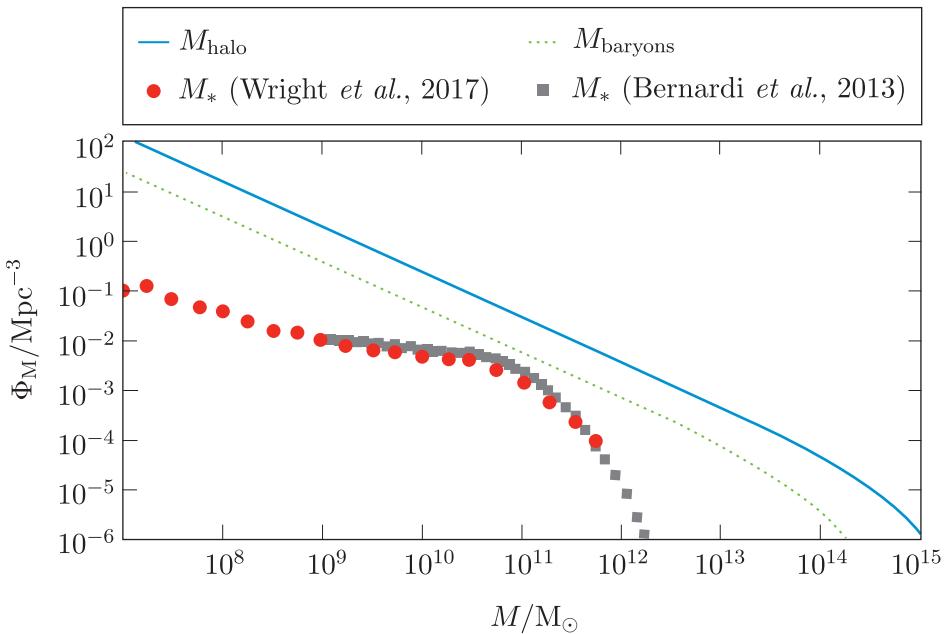


Figure 11.13 The mass function of dark-matter halos (solid blue line). The dotted line is offset to the left by a factor of about six, and shows the baryonic component of those halos. The points show the observed mass of stars in individual galaxies, according to two different surveys.

The shape of the stellar mass function is quite different from that of the dark-matter halos. The halo mass function has a uniform, fairly steep slope across a large mass range (starting to drop off at the very highest masses). The halo mass function in the range $M < 5 \times 10^{13} M_\odot$ can be written as:

$$\frac{dn}{dM_{\text{halo}}} = (1.72 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) M_{\text{halo}}^{-1.9} \quad (11.11)$$

where we have used the derivative with respect to mass rather than logarithmic mass to simplify the analysis.

In contrast, the stellar mass function has a shallow slope at low mass (sometimes known as the **faint-end slope**) and a steep slope at high mass. The turnover point between the two regimes occurs as a characteristic mass M_* corresponding to galaxies of mass fairly similar to the Milky Way.

The next example and exercise explore what can be learned from the different shapes of the halo and stellar mass functions.

Example 11.1

Using Equation 11.11, determine the number density of dark-matter halos in the mass range $10^{10} < M_{\text{halo}} < 10^{11} M_\odot$.

Solution

The differential number of dark-matter halos per cubic megaparsec is given by rearranging Equation 11.11 to give:

$$dn = (1.72 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) M_{\text{halo}}^{-1.9} dM_{\text{halo}}$$

This expression can be integrated to determine the number density of dark-matter halos within a particular mass range, M_1 to M_2 , as follows:

$$\begin{aligned} n &= \int_{M_1}^{M_2} (1.72 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) M_{\text{halo}}^{-1.9} dM_{\text{halo}} \\ &= (1.72 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) \int_{M_1}^{M_2} M_{\text{halo}}^{-1.9} dM_{\text{halo}} \end{aligned}$$

Evaluating the integral gives

$$n = (1.72 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) \left[\frac{M^{-0.9}}{-0.9} \right]_{M_1}^{M_2} \quad (11.12)$$

$$= (1.91 \times 10^8 M_\odot^{0.9} \text{Mpc}^{-3}) [M_1^{-0.9} - M_2^{-0.9}] \quad (11.13)$$

We can now substitute in $M_1 = 10^{10} M_\odot$ and $M_2 = 10^{11} M_\odot$ to give $n = 0.167$ dark-matter halos Mpc^{-3} within the specified mass range.

The next exercise involves some direct comparisons between the halo and stellar mass functions. A useful parameter to introduce at this point is the **star formation efficiency**, ϵ_* . The relationship between a galaxy's stellar mass and the mass of the halo in which it resides depends on ϵ_* as well as the baryon fraction within that halo, f_B (i.e. the fraction of the total halo mass consisting of baryons), according to:

$$M_* = \epsilon_* f_B M_{\text{halo}} \quad (11.14)$$

An efficiency of $\epsilon_* = 1$ would correspond to all the baryons in a dark-matter halo turning into stars.

Exercise 11.6

- (a) Assuming that all dark-matter halos have a universal baryon fraction $f_B = 0.15$ and a star-formation efficiency of $\epsilon_* = 1$, use Equations 11.11 and 11.14 to predict the number density of galaxies in the following stellar mass ranges:
 - (i) $10^8 < M_* < 10^9 M_\odot$
 - (ii) $10^{10} < M_* < 10^{11} M_\odot$
 - (iii) $10^{11.5} < M_* < 10^{12.5} M_\odot$.
- (b) Compare your predictions from part (a) with the *observed* galaxy number densities in the middle of each mass range (Figure 11.13), and comment on what this suggests about the assumptions made.

Exercise 11.6 shows that stars must form much less efficiently in low- and high-mass halos compared to halos with intermediate mass. In fact, both the baryon fraction and star formation efficiency are thought to depend on the halo mass, and this is the explanation for why the shapes of the stellar mass and luminosity function are different from that of the halo mass function.

The processes of stellar and AGN feedback are thought to strongly influence star formation in the least massive and most massive halos. Modern simulations attempt to incorporate these effects to reproduce the observed galaxy luminosity function and other galaxy observations.

11.3.3 Clustering and the two-point correlation function

As well as cataloguing the properties of individual galaxies, the clustering together of galaxies on the sky also provides an important tool for cosmology. The clustering of galaxies can be measured via the angular **two-point correlation function**, $\omega(\theta)$. This is a measure of the distribution of angular separation distances between pairs of galaxies on the sky, here denoted by θ . To obtain this measure, all possible pairs of galaxies are considered, and so it captures whether or not it is more probable for galaxies to be clustered close to another galaxy or not.

To compute $\omega(\theta)$ we can take a galaxy and define a thin annulus (ring) around its location on the sky, with inner and outer angular radii of θ and $\theta + d\theta$, respectively. We use $R(\theta)$ to denote the number of galaxies expected to be found within the annulus if they are randomly distributed – this can be calculated if you know the total number of galaxies in a survey catalogue that is being considered, and the sky area covered. $P(\theta)$ is then defined as the number of galaxies actually observed within that annulus.

The angular correlation function $\omega(\theta)$ is defined as

$$\omega(\theta) = \frac{\langle P(\theta) \rangle}{R(\theta)} - 1 \quad (11.15)$$

where $\langle P(\theta) \rangle$ is obtained by measuring $P(\theta)$ for each galaxy in a survey, and then taking an average.

If the distance to the galaxies is known, then the angular separation, θ , can be converted into a physical separation, r , and a corresponding *physical* two-point correlation function, $\xi(r)$, can be determined. Figure 11.14 shows observations of the galaxy two-point correlation function $\xi(r)$ from a low-redshift sky survey. Here $\xi(r)$ is calculated in a slightly more sophisticated way that accounts for biases that arise from working with different types of galaxy catalogues, but still represents the excess frequency of finding two galaxies at a particular distance from one another compared to a random distribution. Note the steep slope on the left-hand side and the ‘bump’ at around $100h^{-1}$ Mpc, caused by baryon acoustic oscillations, discussed below. (Recall that h represents the Hubble constant, H_0 , in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.)

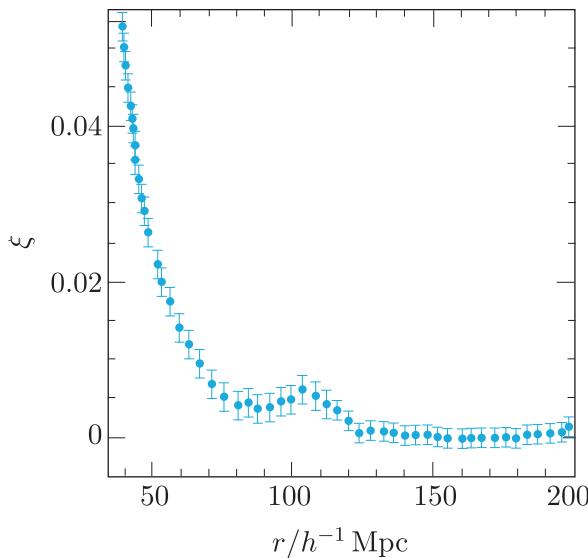


Figure 11.14 Observed galaxy two-point correlation function on large scales.

- What does the steep slope of the two-point correlation function at small r indicate about how galaxies are distributed?
- It is more probable for galaxies to be found at small distances from their neighbours than at large distances. In other words, galaxies cluster together, as predicted from cosmological simulations.

The ‘bump’ feature in Figure 11.14 indicates that there is an excess of galaxy pairs having a separation of $\sim 100h^{-1}$ Mpc (around 150 Mpc for $h = 0.677$). This is caused by a phenomenon known as baryon acoustic oscillations.

11.3.4 Baryon acoustic oscillations

Baryon acoustic oscillations (BAOs) originated in the oscillating photon–baryon fluid prior to recombination. A snapshot of these oscillations in the early Universe was frozen into the distribution of baryonic gas at decoupling. Hence, when the baryons cooled to form galaxies, a signature of these oscillations was imprinted into the distribution of galaxies. Today, the signals associated with BAOs are the largest recognisable structural feature in the Universe.

BAOs had a fixed physical size at the time of decoupling, because of the characteristic speed of sound in the photon–baryon fluid. All acoustic waves in this fluid travelled at a (highly relativistic) speed of $c_s \approx c/\sqrt{3}$. When the radiation pressure supported waves moving outward with respect to the centre of the overdensity dropped suddenly at recombination (see Section 11.1.1), these outward-moving perturbations had all reached the same distance $d_s = c_s t_{\text{rec}}$, known as the sound horizon (see further discussion in Chapter 7). Since the sound horizon is a fixed physical size (determined by the time of photon–baryon decoupling, t_{rec}), observational measurements of the scale on which the BAO signal is found at different

redshifts can be used as a standard ruler to measure the expansion of the Universe. Measurements of the BAO signal have contributed to pinning down the present value of the Hubble parameter, and it is hoped that measuring BAO features at multiple redshifts with future surveys can help to determine the equation of state and nature of dark energy.

Evolution of BAOs

Figure 11.15 illustrates how an initial overdensity of dark matter, baryons and photons evolves in a situation where the photon–baryon fluid is highly overpressured. The horizontal axis shows the *co-moving* size of the perturbation (remember that the true *physical* distance at a given redshift scales with $1 + z$). The vertical axis indicates the mass profile of the perturbation (i.e. the amount of mass for each type of material at a given radius). We can think of the mass profile as the amplitude of the perturbation as it propagates out. Note that the values on the vertical axis increase from one panel to the next as time increases. This is because the perturbation grows and the dark-matter halo collapses.

We can work through Figure 11.15 to explore how the different particle types interact and hence how the BAO signal arose. First, consider a dark-matter overdensity (i.e. density perturbation) into which more dark matter falls under gravity. The photon–baryon fluid also falls into this overdensity, but radiation pressure causes the fluid to rebound, which drives an acoustic wave outwards at the speed of sound.

- Panel (a) of Figure 11.15 shows the evolution of the density perturbation considered above shortly after its formation. Interpret the lines in this panel.
- The dark matter is starting to accumulate at the centre of the perturbation. Most of the mass is concentrated within about 10 Mpc co-moving distance (corresponding to 1.5 kpc physical distance at redshift $z = 6824$). Meanwhile, the photons and baryons, which are tightly coupled together at this redshift, are already expanding well beyond the concentration of dark matter, reaching a co-moving distance of 30 Mpc.

As we move to (b), dark matter is continuing to fall into the perturbation from larger radii. However, particle energy stops the dark matter from collapsing further, so it remains as a diffuse, overdense blob with a co-moving size of ~ 10 Mpc. The perturbation in the photon–baryon fluid continues to propagate out at the speed of sound.

Between panels (b) and (c), last scattering occurs. We can see the photons keep expanding out in panels (c) and (d). However, the baryons remain almost fixed in place, concentrated in a spherical shell at a co-moving distance of ~ 150 Mpc from the initial perturbation, which corresponds to a physical scale of ≈ 140 kpc at recombination, i.e. $z \approx 1100$.

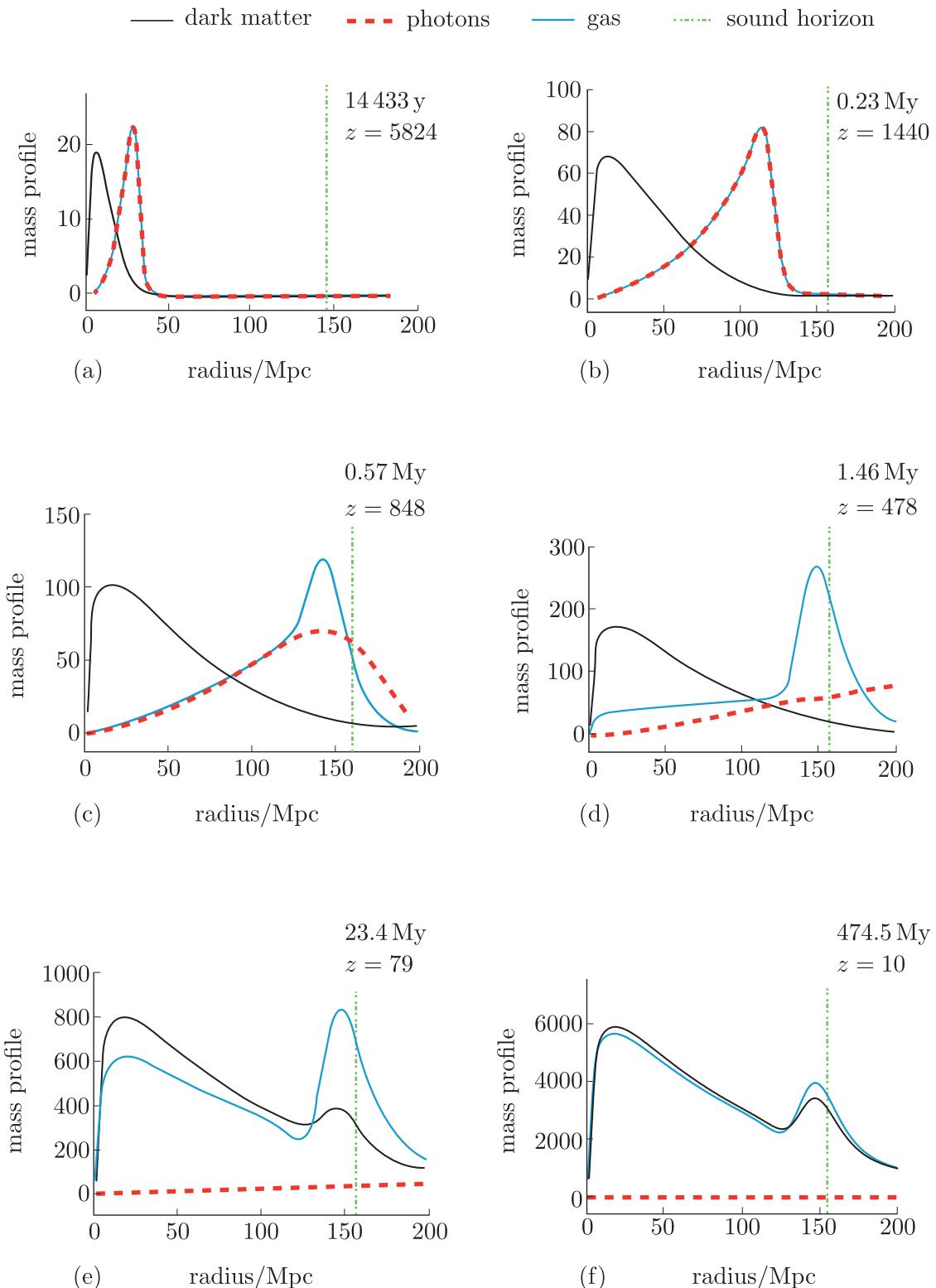


Figure 11.15 The evolution of a single, initially highly overpressured perturbation in the photon–baryon fluid, as well as the corresponding dark-matter perturbation, at high redshift. The horizontal axis shows *comoving* distance, and so reflects the *present-day* size of the regions considered.

Between panels (b) and (c), the dominant force on baryons stops being the electromagnetic interaction with photons, and starts becoming the gravitational attraction to the dark matter. Therefore, baryons begin falling into the nearest dark-matter halo. However, because baryons make up a significant fraction ($\approx 15\%$) of matter, this also has a gravitational effect on the dark matter, and so the dark-matter perturbation starts to expand a little towards the baryonic perturbation as well, as becomes more evident in panel (d).

By panel (d), about a million years after recombination and last scattering, we can see these collapse processes at work: the density of baryonic matter is now increasing in the inner regions. This continues through panels (e) and (f), until the mass profile amplitudes of the inner dark-matter and baryonic overdensities are almost equal.

We can also see from panels (d)–(f) that the dark matter is beginning to cluster on larger scales around the baryonic matter peak at 150 Mpc. Hence, by panel (f), at $z = 10$, dark matter and baryonic matter have become highly coupled. This process gives us two peaks in the matter profile: one peak around ~ 10 Mpc seeded the most massive halos we see today; and a smaller peak at 150 Mpc, which is the BAO signature and represents a characteristic scale on which galaxies tend to be separated. The observable consequence of this smaller, outer peak is the bump feature in the two-point correlation function seen in Figure 11.14.

11.4 Summary of Chapter 11

- Before recombination, baryons were physically coupled to radiation, and the radiation pressure in the photon–baryon fluid was high. After decoupling, the baryonic gas collapsed into nearby dark-matter overdensities. These overdensities continued to grow during a period known as the dark ages, from $z \approx 1100$, when the Universe was $\sim 380\,000$ years old and lasting for around 200 million years, until $z \approx 20$.
- Self-interaction among baryons allows them to cool in a way dark matter cannot, so they can collapse into smaller clumps, including individual stars. The **cooling timescale** t_{cool} depends on gas density, temperature and the available cooling mechanisms, characterised by a **cooling rate** Λ_{cool} according to

$$t_{\text{cool}} = \frac{\gamma n k_B T}{2\Lambda_{\text{cool}}} \quad (\text{Eqn 11.1})$$

where γ is the number of degrees of freedom of the dominant particle type.

- The first (**Population III**) stars probably formed around $z \approx 17$ and were likely very massive. They ionised **Strömgren spheres** in the interstellar medium (ISM) around them.

- The size of a Strömgren sphere is limited by recombination to a radius of

$$R_{\max} \approx \left[\frac{3Q}{4\pi n_i n_e \alpha(T)} \right]^{1/3} \quad (\text{Eqn 11.5})$$

- The sphere's expansion slows exponentially as recombination becomes more significant, with expansion characterised approximately by

$$R(t) \approx R_{\max} [1 - \exp(-t/t_r)]^{1/3} \quad (\text{Eqn 11.6})$$

- The dense ISM within galaxies can recombine, allowing star formation to continue. However, recombination was extremely rare in intergalactic space at that time. The merged Strömgren spheres from entire galaxies rapidly ionised the diffuse **intergalactic medium** (IGM), probably by $z \approx 6$. It remains ionised to this day.
- Matter cycles in and out of stars over time. Stellar winds and supernovae release material back to the ISM, but some remains locked in compact **stellar remnants**.
- The nuclear-processed matter ejected from stars causes chemical enrichment of the ISM, spreading out astronomical metals within it and increasing the metallicity of the host galaxy.
- Even when stars die, explosions and mergers of compact objects can be sources of additional chemistry and feedback via supernovae and other outflows.
- Supermassive black holes** (SMBHs) exist in the centre of galaxies. These may form by mergers of smaller compact objects through the process of **mass segregation**. SMBHs can form **active galactic nuclei** (AGN) and their associated **jets** and **radio lobes**.
- Stellar feedback** and **AGN feedback** can both increase or decrease the star-formation rate of a galaxy, potentially driving outflows from the galaxy's gas reservoir.
- Galaxy **number counts** can be used to investigate the geometry of the Universe, while **luminosity functions** and **stellar mass functions** provide sophisticated tests of theoretical models and computer simulations that predict the evolution of structure in the Universe.
- The mass function for dark-matter halos can be modelled as a power law below the scale of galaxy clusters, while the stellar mass function has a more complicated form due to a varying efficiency of star formation with halo mass.
- The galaxy **two-point correlation function** measures the clustering of galaxies as a function of separation distance, and is a useful cosmological probe.
- Baryon acoustic oscillations** (BAOs) are acoustic waves in the early Universe that imprint a 'bump' in the two-point correlation function at a characteristic size scale determined by the sound horizon at recombination. This scale is used as a standard ruler in cosmology.

Solutions to exercises

Solution to Exercise 6.1

Using the fact that the real Universe is spatially flat, the proper distance between two points at time t is directly proportional to the scale factor $a(t) = (1+z)a(t_0) = 1+z$, so a region with proper size $l = 19 \text{ kpc}$ at $z_{\text{ls}} \approx 1090$ has proper size $\Delta l(1+z_{\text{ls}}) \approx 21 \text{ Mpc}$ today.

Solution to Exercise 7.1

Start by using Equation 3.15 from Chapter 3 to calculate the ratio of the scale factors corresponding to z_{ls} and z_{mr} :

$$\frac{a_{\text{ls}}}{a_{\text{mr}}} = \frac{1+z_{\text{mr}}}{1+z_{\text{ls}}} = \frac{5731}{1091} \approx 5.25 \quad (\text{S8})$$

Now we can use Equation 7.2 to calculate the ratio of ages of the model universe when $a = a_{\text{mr}}$ and when $a = a_{\text{ls}}$. Note that the coefficient *outside* the square brackets in Equation 7.2 will cancel in the ratio.

When $a = a_{\text{mr}}$ the term *inside* the square brackets reduces to $1 - \sqrt{2}/2$, and we can use the result from Equation S8 to write

$$\begin{aligned} \frac{t_{\text{mr}}}{t_{\text{ls}}} &= \frac{\left(1 - \frac{\sqrt{2}}{2}\right)}{1 - \left(1 - \frac{a_{\text{ls}}}{2a_{\text{mr}}}\right) \left(1 + \frac{a_{\text{ls}}}{a_{\text{mr}}}\right)^{1/2}} \\ &= \frac{\left(1 - \frac{\sqrt{2}}{2}\right)}{1 - \left(1 - \frac{5.25}{2}\right) (1 + 5.25)^{1/2}} \\ &= 0.0579 \end{aligned}$$

We have therefore shown that this universe had only been radiation-dominated for $\sim 5.8\%$ of its history by the epoch of last scattering. For the subsequent $\sim 94\%$ of its history, up until the time when the CMB photons were released, energy density had been dominated by matter, of which dark matter formed the dominant component. You may have reached the same conclusion by

separately calculating $t(a_{\text{ls}})$ and $t(a_{\text{mr}})$ and comparing their values.

Solution to Exercise 7.2

(a) The question tells us to assume that $\Omega_r \gg \Omega_b$. This means that $\Omega_b/\Omega_r \ll 1$, and so the expression for the speed of sound in Equation 7.5 can be simplified to

$$c_s = c \left(3 + \frac{9}{4} \frac{\Omega_b}{\Omega_r} \right)^{-1/2} \approx \frac{c}{\sqrt{3}}$$

Substituting this approximation in Equation 7.4 and using the expression for the horizon distance in Equation 7.1 yields the required expression for the maximum distance an acoustic oscillation could have travelled by the time of last scattering:

$$\begin{aligned} d_s(z_{\text{ls}}) &= a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{c_s}{a(t)} dt \\ &\approx a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{c}{\sqrt{3} a(t)} dt \\ &\approx \frac{c}{\sqrt{3}} a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{dt}{a(t)} \\ &\approx \frac{d_{\text{hor}}(z_{\text{ls}})}{\sqrt{3}} \end{aligned}$$

(b) Now we can calculate the acoustic scale using Equation 7.6:

$$\begin{aligned} \theta_s &\approx \frac{d_s(z_{\text{ls}})}{d_A(z_{\text{ls}})} \approx \frac{d_{\text{hor}}(z_{\text{ls}})}{\sqrt{3} d_A(z_{\text{ls}})} \\ &\approx \frac{0.29 \text{ Mpc}}{\sqrt{3} \times 12.73 \text{ Mpc}} = 0.013 \text{ rad} \approx 0.75^\circ \end{aligned}$$

As you learned in Section 6.3.1, fluctuations with angular scale $\Delta\theta$ radians correspond to multipole number $\sim \pi/\ell$. Therefore, because π radians is equivalent to 180 degrees, the angular scale represented by θ_s corresponds to multipole number

$$\ell_s \approx \frac{180^\circ}{0.75^\circ} \approx 240$$

which is remarkably close to the observed location of the first acoustic peak in Figure 7.1.

Solution to Exercise 8.1

Your completed table should look similar to Table S2. As noted in the question, you may have found somewhat different numbers for the relative strengths of the forces. Those quoted here are the relative strength of the four forces acting between two protons that are just in contact.

Table S2 Summary of the basic properties of the four forces.

Interaction	Particles affected	Carrier particle	Relative strength
Strong	quarks	gluon	1
Weak	quarks, baryons, leptons	W and Z bosons	10^{-2}
Electromagnetic	all charged particles, leptons	photon	10^{-7}
Gravity	all massive particles	none (Standard Model)	10^{-39}

Solution to Exercise 8.2

The interaction energy in J is

$$E_{\text{int}} \approx 1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^{14} \text{ K}$$

$$\approx 1.381 \times 10^{-9} \text{ J}$$

Converting to GeV:

$$\frac{1.381 \times 10^{-9} \text{ J}}{1.602 \times 10^{-19} \text{ J eV}^{-1}} = 8.620 \times 10^9 \text{ eV}$$

$$\approx 8.6 \text{ GeV}$$

Solution to Exercise 8.3

As you read in Section 8.1.1, in order to create particles the interaction energy must be higher than the total rest mass energy of the corresponding particles, as given by $E = mc^2$. Therefore:

- a proton–antiproton pair requires energy of $E = 2m_p c^2 = 3.0 \times 10^{-10} \text{ J} = 1.9 \text{ GeV}$
- an electron–positron pair requires $E = 2m_e c^2 = 1.6 \times 10^{-13} \text{ J} = 0.001 \text{ GeV}$.

So, in terms of the interaction energy required, it is much easier to create electron–positron pairs than proton–antiproton pairs.

Solution to Exercise 8.4

Using Equation 8.15 and substituting in the relevant particle values, converting to appropriate units and recalling that the interaction energy E is approximately equal to $k_B T$ (Equation 8.2) gives:

$$\begin{aligned} \frac{n_n}{n_p} &= \exp \left[-\frac{(m_n - m_p)c^2}{k_B T} \right] \\ &= \exp \left(-\frac{2.067 \times 10^{-13} \text{ kg m}^2 \text{ s}^{-2}}{1.282 \times 10^{-13} \text{ kg m}^2 \text{ s}^{-2}} \right) \\ &= 0.20 \end{aligned}$$

In other words, there are five protons for every neutron, as required.

Solution to Exercise 8.5

We can rearrange Equation 8.16 to find an expression for t :

$$\frac{N(t)}{N(t_{\text{init}})} = \exp(-t/\tau_n)$$

and so

$$t = -\tau_n \ln[N(t)/N(t_{\text{init}})]$$

Because we are told that n_p can be approximated as remaining constant (because $n_p \gg n_n$ when the decay starts), the ratio of the number of neutrons is the same as the ratio between n_n/n_p at the two times:

$$\frac{N(t)}{N(t_{\text{init}})} = \frac{n_n/n_p(t)}{n_n/n_p(t_{\text{init}})}$$

Therefore t can be obtained from

$$t = -\tau_n \ln \left(\frac{n_n/n_p(t)}{n_n/n_p(t_{\text{init}})} \right)$$

Substituting in the provided values gives $t = 1230 \text{ s}$. Because t is the time elapsed since $t_{\text{init}} = 1 \text{ s}$, a more precise estimate of the time since the big bang at which this neutron-to-proton ratio would apply is 1231 s.

Solution to Exercise 8.6

Please refer to the Exercise 8.6 Jupyter Notebook solution available from the online module resources to see the solution to this example.

The most likely pitfall when attempting this task is in setting the integration ranges – this requires converting from photon energy to frequency, and using the provided hint to obtain a maximum frequency for the calculation of the total number of photons.

Solution to Exercise 9.1

We can make use of the final equation obtained in Example 9.1 (Equation 9.5) to answer this question. It is not trivial to solve this expression algebraically, but we can simply substitute values into the left- and right-hand sides and demonstrate equality.

Taking the definition of recombination corresponding to $X = 0.5$, the left-hand side of Equation 9.5 becomes:

$$\frac{1-X}{X^2} = 2.0$$

Substituting the provided temperature and value of η into the right-hand side gives:

$$10^{-9} \left[\frac{a(3795)^3}{3k_B} \right] \times \left[\frac{m_e k_B (3795)}{2\pi\hbar^2} \right]^{-3/2} \times \exp \left[\frac{Q}{k_B (3795)} \right] = 2.00$$

Therefore the left- and right-hand sides are in agreement (to within a few per cent) for $T = 3795$ K.

Solution to Exercise 9.2

The two measured D/H values of 2.52×10^{-5} (quasars) and 1.38×10^{-5} (Milky Way), correspond to the abundances predicted by big bang nucleosynthesis for $\eta \sim 6 \times 10^{-10}$ and $\sim 9 \times 10^{-10}$, respectively, according to Figure 9.3. (Note that these are necessarily fairly rough estimates given the precision of the graph.)

We can convert these to values of $\Omega_{b,0}$ using Equation 8.11 from Chapter 8, which rearranges

to give:

$$\Omega_{b,0} = \eta \frac{\langle m \rangle n_{\gamma,0}}{\rho_{c,0}}$$

If $\langle m \rangle = 1.6726 \times 10^{-27}$ kg, $n_{\gamma,0} = 4.0 \times 10^8$ m⁻³ and $\rho_{c,0} = 8.599 \times 10^{-27}$ kg m⁻³, this gives $\Omega_{b,0} \approx 0.05$ (quasars) and $\Omega_{b,0} \approx 0.07$ (Milky Way). Hence the quasar-measured estimate of η is in much better agreement with the baryon density parameter measured via the CMB (given as 0.0490 in the table of constants) than the Milky Way estimate.

Solution to Exercise 9.3

If the baryon-to-photon ratio is low then the start of deuterium production is delayed, because the high-energy tail of the photon distribution is more successful at dissociating the deuterons as they form. The consequence of this deuterium bottleneck is that the ratio of neutrons to protons continues to decrease, so there are fewer neutrons available for nucleosynthesis when η is low compared to when it is high.

The vast majority of deuterium produced in the BBN period goes on to form helium via the reactions shown in Section 8.3.2, so more helium forms for high η (when lots of neutrons are available) than for low η . The decreasing abundance of deuterium occurs because the reactions that create helium are more efficient when there are more neutrons present, so a greater proportion of the deuterium is converted to helium.

Solution to Exercise 9.4

(a) The total mass can be calculated by substituting the provided velocity dispersion and radius into Equation 9.18, which gives $M = 4.84 \times 10^{45}$ kg = 2.43×10^{15} M_⊕.

(b) The mass-to-light ratio is simply the total mass, in units of solar mass, divided by the optical luminosity, in units of solar luminosity. In other words $M/L = (2.43 \times 10^{15} \text{ M}_{\odot}) / (5.0 \times 10^{12}) \approx 490$.

(c) By definition, the Sun has a mass-to-light ratio of 1 when considered in units of M_⊕ and L_⊕. Therefore, a system with $M/L \gg 1$ has proportionally much more mass for the same

Solutions to exercises

amount of light it is producing. If the Coma cluster's mass was primarily made up of stars, then on average the stars would need to be >500 times less luminous than the Sun. This is not realistic.

Solution to Exercise 9.5

Rearranging Equation 9.20 for mass gives:

$$M = \frac{5k_BRT}{G\bar{m}}$$

Substituting temperatures of 10^7 K and 10^8 K gives a mass range of 3.2×10^{44} – 3.2×10^{45} kg, which corresponds to 1.6×10^{14} – 1.6×10^{15} M_⊕.

Solution to Exercise 10.1

The gravitational acceleration of the two galaxies would be:

$$\begin{aligned} g &= \frac{GM}{r^2} = \frac{G \times 10^{12} \text{ M}_\odot}{(3 \text{ Mpc})^2} \\ &= \frac{6.673 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2} \times 1.99 \times 10^{42} \text{ kg}}{(3 \times 3.086 \times 10^{22} \text{ m})^2} \\ &= 1.55 \times 10^{-14} \text{ m s}^{-2} \end{aligned}$$

It would therefore take

$$\frac{200 \text{ km s}^{-1}}{1.55 \times 10^{-14} \text{ m s}^{-2}} = 1.29 \times 10^{19} \text{ s}$$

or about 30 times the current age of the Universe to bring them to rest. Given that H_0 appears to be increasing (i.e. the expansion is accelerating), it seems these two galaxies will never come together.

Solution to Exercise 10.2

The mean density for each object can be obtained by assuming a sphere of radius, R , so that

$$\rho = \frac{M}{\frac{4}{3}\pi R^3}$$

The free-fall timescale can then be obtained via Equation 10.6:

$$t_{\text{ff}} \approx 18.45 \text{ hours} \left(\frac{\rho}{1 \text{ kg m}^{-3}} \right)^{-1/2}$$

The resulting values for each object are given in Table S3.

Table S3 Densities and free-fall timescales for the Milky Way and the Local Group.

System	$\rho/\text{kg m}^{-3}$	t_{ff}
Milky Way	1.4×10^{-21}	57 My
Local Group	6.0×10^{-26}	8.6 Gy

These results show that objects of the density of the present-day Milky Way can collapse on a relatively short cosmological timescale if unopposed: 57 million years is less than the time since the demise of the dinosaurs! By contrast, a halo of the mean density of the Local Group would require a substantial fraction (about half) of the age of the Universe to collapse.

This calculation started from the densities of the present-day (collapsed) objects. The collapse timescales to form these objects starting from the less dense conditions of a primordial cloud would, in principle, be longer.

Solution to Exercise 10.3

Equation 10.16 states $\delta(t) \propto 1/(1+z)$.

The question states that at the time of decoupling, $z \approx 1090$, so $1/(1+z) \approx 9.17 \times 10^{-4}$.

At $z \approx 13$, we have $1/(1+z) \approx 7.14 \times 10^{-2}$.

If $\delta \approx 10^{-5}$ at decoupling, as implied by fluctuations in the CMB temperature, then δ will have only grown to $(10^5) \times (7.14 \times 10^{-2}/9.17 \times 10^{-4}) \sim 8 \times 10^{-4}$ by $z \approx 13$, and $\delta \approx 10^{-2}$ today. Galaxies have densities many times higher than the mean density of the Universe, i.e. $\delta > 1$. Although Equation 10.16 is invalid as δ approaches 1, this exercise shows it should not yet have done so. In other words, Equation 10.16 implies that galaxies should not exist today, never mind at $z \approx 13$.

The solution to this apparent inconsistency is to recall that the density fluctuations in the CMB describe the behaviour of the photon–baryon fluid, not the dark matter. In order for galaxy-sized overdensities to form, the dark-matter overdensities at the time the CMB was produced must have been larger by a substantial factor than those of the baryons and photons.

Solution to Exercise 11.1

- (a) The gas is able to cool over the density range where the track has a negative slope (i.e. temperature is decreasing as density increases); this occurs between the edge of the diagram at $n = 1 \text{ cm}^{-3}$ and $n \approx 10^5 \text{ cm}^{-3}$. Over the rest of the density range (i.e. where $n > 10^5 \text{ cm}^{-3}$) the gas heats as it becomes denser, and therefore cannot fragment into smaller clumps.
- (b) The solar-composition line reaches a minimum just to the right of the line corresponding to a Jeans mass of $1 M_\odot$. This is the optimal place for star formation, as it has the highest density and lowest temperature. It suggests that typical stars would have masses a little less than that of the Sun, which is indeed what is observed in the Milky Way.
- (c) Considering the $Z = 0$ cooling track, which corresponds to the expected behaviour of primordial gas, the minimum occurs at a much higher Jeans mass than the solar-composition line ($> 100 M_\odot$). This is because the cooling mechanisms involving emission lines are very much less effective at low metallicity. The first generation of stars were therefore probably much more massive than stars today, with an average mass of several hundred solar masses.

Solution to Exercise 11.2

- (a) The cooling timescale can be estimated from Figure 11.3. For a gas cloud with $Z = 0$ and $T \approx 3000 \text{ K}$ and no molecular hydrogen, Figure 11.3 indicates that $\Lambda_{\text{cool}} \approx 3 \times 10^{-29} \text{ J m}^{-3} \text{ s}^{-1}$. Using this value in Equation 11.1 with $\gamma = 3$ (assuming monatomic hydrogen) gives:

$$\begin{aligned} t_{\text{cool}} &= \frac{3nk_B T}{2\Lambda_{\text{cool}}} \\ &\approx \frac{3 \times 3 \times 10^8 \text{ m}^{-3} \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 3000 \text{ K}}{2 \times 3 \times 10^{-29} \text{ J m}^{-3} \text{ s}^{-1}} \\ &\approx 6.2 \times 10^{17} \text{ s} \\ &\approx 20 \text{ Gy} \end{aligned}$$

We have assumed that the gas started cooling from 3000 K at around the time of recombination, t_{rec} ,

and so the age of the Universe at which stars first formed is $t_{\text{cool}} + t_{\text{rec}}$. In this case the timescale is dominated by the cooling time, and so the age of the Universe when stars first formed would still be 20 Gy.

- (b) For a gas cloud with 1% molecular hydrogen, when $T \approx 3000 \text{ K}$ then $\Lambda_{\text{cool}} \approx 3 \times 10^{-25} \text{ J m}^{-3} \text{ s}^{-1}$. Using this value in Equation 11.1, this time with $\gamma = 6$ for a molecular gas, gives:

$$\begin{aligned} t_{\text{cool}} &= \frac{6nk_B T}{2\Lambda_{\text{cool}}} \\ &\approx \frac{6 \times 3 \times 10^8 \text{ m}^{-3} \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 3000 \text{ K}}{2 \times 3 \times 10^{-25} \text{ J m}^{-3} \text{ s}^{-1}} \\ &\approx 1.2 \times 10^{14} \text{ s} \\ &\approx 4 \text{ My} \end{aligned}$$

In this case the age of the Universe at which stars first formed is $t_{\text{cool}} + t_{\text{rec}} \approx 4.4 \text{ My}$.

- (c) The age of the Universe when stars first formed calculated in part (a) for the zero metallicity case is considerably greater than the actual age of the Universe according to current models and observations ($\approx 13.7 \text{ Gy}$). This provides strong evidence that molecular hydrogen cooling must be important to enable the first stars to form: without it there would not yet be stars, and we wouldn't be here to think about them. The age calculated in part (b) for the 1% H_2 scenario is quite short: in reality it is thought that the first stars took around 200 million years to form, and so although we have shown that molecular hydrogen must be important, more sophisticated modelling is needed to obtain a more accurate timescale for the first stars.

- (d) The main simplification that has been made in our calculations is to ignore the fact that the parameters of Equation 11.1 evolve:

- the baryons initially continue to expand and dilute with the Universe, so n goes down
- as a cloud of baryons collapses into a star, n goes up
- the gas is still being heated by the CMB, so T stays high
- the cooling rate, Λ_{cool} , will change as the gas cools (and as it becomes denser).

Solutions to exercises

Accounting for the evolution of these parameters, and more accurate modelling of the H₂ fraction, would enable a more accurate timescale to be obtained.

Solution to Exercise 11.3

(a) We first note that

$$\begin{aligned} 10 \text{ cm}^{-3} &= 10 \times (10^{-2} \text{ m})^{-3} \\ &= 10^7 \text{ m}^{-3} \end{aligned}$$

Then for scenario (i) the maximum radius is:

$$\begin{aligned} R_{\max} &\approx \left[\frac{3Q}{4\pi n_i n_e \alpha(T)} \right]^{1/3} \\ &\approx \left[\frac{3 \times 10^{50} \text{ s}^{-1}}{4\pi \times (10^7 \text{ m}^{-3})^2 \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1}} \right]^{1/3} \\ &\approx 8.42 \times 10^{17} \text{ m} \\ &\approx 27 \text{ pc} \end{aligned}$$

For scenario (ii) first note that:

$$\begin{aligned} 10^{-5} \text{ cm}^{-3} &= 10^{-6} \times (10^{-2} \text{ cm})^{-3} \\ &= 10 \text{ m}^{-3} \end{aligned}$$

Then the maximum radius is:

$$\begin{aligned} R_{\max} &\approx \left[\frac{3 \times 10^{56} \text{ s}^{-1}}{4\pi \times (10 \text{ m}^{-3})^2 \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1}} \right]^{1/3} \\ &\approx 2.71 \times 10^{23} \text{ m} \\ &\approx 8.8 \text{ Mpc} \end{aligned}$$

(b) For scenario (i) the recombination timescale is:

$$\begin{aligned} t_r &= (10^7 \text{ m}^{-3} \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1})^{-1} \\ &= 2.5 \times 10^{11} \text{ s} \\ &= 8000 \text{ y} \end{aligned}$$

while for scenario (ii):

$$\begin{aligned} t_r &= (10 \text{ m}^{-3} \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1})^{-1} \\ &= 2.5 \times 10^{17} \text{ s} \\ &\approx 7.9 \times 10^9 \text{ y} \end{aligned}$$

(c) For scenario (i) the size of the Strömgren sphere after 10⁶ years is:

$$\begin{aligned} R(t) &\approx R_{\max} [1 - \exp(-t/t_r)]^{1/3} \\ &\approx 27 \text{ pc} [1 - \exp(-10^6/8000)]^{1/3} \\ &\approx 27 \text{ pc} \end{aligned}$$

and for scenario (ii):

$$\begin{aligned} R(t) &\approx 8.8 \text{ Mpc} [1 - \exp(-10^6/(7.9 \times 10^9))]^{1/3} \\ &\approx 0.44 \text{ Mpc} \end{aligned}$$

[For interest, this is around half the distance light can travel in that time. Consequently, we expect the Strömgren sphere to expand out of this (small) galaxy at an appreciable fraction of the speed of light.]

Solution to Exercise 11.4

(a) Using the numbers given, the escape speed is

$$\begin{aligned} v_{\text{esc}} &\approx \sqrt{\frac{2G \times 10^{12} \text{ M}_\odot}{8 \text{ kpc}}} \\ &\approx 1000 \text{ km s}^{-1} \end{aligned}$$

(b) The escape speed is (i) slightly greater than that of the supernova, but (ii) much greater than that of the low-mass star.

(c) Supernovae ejecta will therefore not escape from the galaxy, but will be flung out far into the galaxy's halo before falling back down. Ejecta from dying low-mass stars will contribute directly to its local environment, cycling matter directly back into the galaxy's interstellar medium.

(d) For a galaxy of $M = 10^8 \text{ M}_\odot$ and $R = 800 \text{ pc}$, the escape speed will only be $v_{\text{esc}} \approx 33 \text{ km s}^{-1}$. Consequently, while the galaxy will still retain the ejecta of its low-mass stars, it will lose the ejecta of its supernovae. This will lead to a different overall chemical composition for the smaller galaxy as its retained ejecta will be enriched only in the elements created by dying low-mass stars, which can be seen in Figure 11.8 to differ from those produced in supernovae (exploding massive stars).

Solution to Exercise 11.5

- (a) As indicated in Figure 11.8, hydrogen is formed in the big bang, oxygen is formed from exploding massive stars, and (although some iron is produced by exploding massive stars) most iron is produced by exploding white dwarfs.
- (b) To begin with, the Universe is composed of hydrogen and helium. Recalling that time evolution proceeds from left to right in Figure 11.7, massive stars will be the first to die. However, for white dwarfs to reach the point where they inject metals into the ISM, less-massive stars have to go through their entire (longer) evolution to produce white dwarfs. The white dwarfs then have to evolve further until they become Type Ia supernovae.
- (c) As massive stars die, the oxygen and iron abundances will increase in tandem, creating a constant [O/Fe] ratio and increasing [Fe/H] ratio. As white dwarfs explode and contribute more iron, the [Fe/H] ratio will continue to increase, but the extra iron will make the [O/Fe] ratio decrease.
- (d) The different points where the [O/Fe] ratio decreases mark the different levels of enrichment each galaxy achieved by the time exploding white dwarfs started to contribute. (While the first white dwarfs form from intermediate-mass stars in 35 My, the process of chemical enrichment takes much longer, as they then need to accrete enough matter to become Type Ia supernovae.)

Solution to Exercise 11.6

- (a) In order to predict the number density of galaxies in a particular stellar mass range we first need to identify the corresponding halo mass range for each case. The question tells us to assume $f_B = 0.15$ and $\epsilon_* = 1$. We can therefore assume that all of the baryons in the corresponding halos are now in the form of stars, and so the total mass of baryons in each halo is the same as the stellar mass. This means that, according to Equation 11.14, the corresponding halo mass range must be $1/0.15 \approx 6.7$ times the stellar mass range.

The relevant mass ranges to consider are therefore:

- (i) $6.7 \times 10^8 < M_{\text{halo}} < 6.7 \times 10^9 M_\odot$
- (ii) $6.7 \times 10^{10} < M_{\text{halo}} < 6.7 \times 10^{11} M_\odot$
- (iii) $2.1 \times 10^{12} < M_{\text{halo}} < 2.1 \times 10^{13} M_\odot$.

We can now determine the number density of halos using the same method as for Example 11.1. Substituting each of the mass ranges into Equation 11.13 gives galaxy densities of

- (i) $n = 1.9 \text{ galaxies Mpc}^{-3}$
- (ii) $n = 0.03 \text{ galaxies Mpc}^{-3}$
- (iii) $n = 0.001 \text{ galaxies Mpc}^{-3}$.

(b) Estimating the observed galaxy number densities for the middle of each *stellar mass range* from Figure 11.13 gives a little above ~ 0.03 , ~ 0.01 and $\sim 10^{-5}$, respectively. Therefore, the assumptions of constant baryon fraction and star-formation efficiency cannot be correct. The prediction is closest for the intermediate mass range, but predicts far too many galaxies for both the low and high stellar mass ranges.

References and acknowledgements

References

- Bernardi, M. *et al.* (2013) ‘The massive end of the luminosity and stellar mass functions: dependence on the fit to the light profile’, *Monthly Notices of the Royal Astronomical Society*, 436(1), pp. 697–704. Available at <https://doi.org/10.1093/mnras/stt1607>.
- Cooke, R. J. *et al.* (2014) ‘Precision measurements of the primordial abundance of deuterium’, *The Astrophysical Journal*, 781(1), 1. Available at <https://doi.org/10.1088/0004-637X/781/1/31>.
- Cyburt, R. H. *et al.* (2016) ‘Big bang nucleosynthesis: present status’, *Reviews of Modern Physics*, 88(1), 015004. Available at <https://doi.org/10.1103/RevModPhys.88.015004>.
- Lellouch, E. *et al.* (2001) ‘The deuterium abundance in Jupiter and Saturn from ISO-SWS observations’, *Astronomy & Astrophysics*, 370(2), pp. 610–622. Available at <https://doi.org/10.1051/0004-6361:20010259>.
- Linsky, J. L. *et al.* (2006) ‘What is the total deuterium abundance in the local Galactic disk?’, *The Astrophysical Journal*, 647(2), pp. 1106–1124. Available at <https://doi.org/10.1086/505556>.
- Ryden, B. (2017) *Introduction to cosmology*. 2nd edn. New York: Cambridge University Press.
- Sbordone, L. *et al.* (2010) ‘The metal-poor end of the Spite plateau’, *Astronomy & Astrophysics*, 522, A26. Available at <https://doi.org/10.1051/0004-6361/200913282>.
- Wright, A. H. *et al.* (2017) ‘Galaxy And Mass Assembly (GAMA): the galaxy stellar mass function to $z = 0.1$ from the r -band selected equatorial regions’, *Monthly Notices of the Royal Astronomical Society*, 470(1), pp. 283–302. Available at <https://doi.org/10.1093/mnras/stx1149>.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: Max-Planck Institute for Physics.

Chapter images: Figure 6.2a: BICEP/Keck Collaboration; Figure 6.2b: Amble, https://commons.wikimedia.org/wiki/File:South_pole_telescope_nov2009.jpg, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported (CC BY SA 3.0) license, <https://creativecommons.org/licenses/by-sa/3.0/deed.en>; Figure 6.2c: Nathan Precup; Figure 6.3: NASA/JPL-Caltech/ESA; Figures 6.5, 6.7 and 6.9: this research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology; Figure 6.8: ESO, <https://arxiv.org/abs/1502.01588>, licensed under a Creative Commons Attribution 4.0 International (CC BY SA 3.0) license, <https://creativecommons.org/>

References and acknowledgements

licenses/by/4.0/; Figure 7.12: adapted from Suzuki, N. *et al.* (2012) ‘The Hubble Space Telescope cluster supernova survey. V. Improving the dark-energy constraints above $z > 1$ and building an early-type-hosted supernova sample’, *The Astrophysical Journal*, 746(1), p. 85, The American Astronomical Society; Figure 9.2: Sloan Digital Sky Survey; Figure 9.6: Izotov, Y. I. *et al.* (2014) ‘A new determination of the primordial He abundance using the HeI $\lambda 10830 \text{ \AA}$ emission line: cosmological implications’, *Monthly Notices of the Royal Astronomical Society*, 445(1), pp. 778–793, Royal Astronomical Society; Figure 9.7: adapted from Figure 15 of Sbordone, L., *et al.* (2010) ‘The metal-poor end of the Spite plateau’, *Astronomy & Astrophysics* 522, A26, EDP Sciences, https://www.aanda.org/articles/aa/full_html/2010/14/aa13282-09/F15.html; Figure 9.9a: NASA/STScI; Figure 9.9b: NASA/CXC/MIT/Peng, E.-H. *et al.*; Figure 9.10: NASA/STScI, Magellan/U. Arizona/Clowe, D. *et al.*; Figure 9.11a: Justin Yaros and Andy Schleif/Flynn Haase/NOAO/AURA/NSF; Figure 9.11b: adapted from Begeman, K. G. *et al.* (1991) ‘Extended rotation curves of spiral galaxies: dark haloes and modified dynamics’, *Monthly Notices of the Royal Astronomical Society*, 249(3), pp. 523–537; Figure 9.12: Sofue, Y. *et al.* (2009) ‘Unified rotation curve of the Galaxy – decomposition into de Vaucouleurs bulge, disk, dark halo, and the 9-kpc rotation dip’, *Astronomical Society of Japan*, 61(2), Astronomical Society of Japan; Figure 10.2a: Bender, R. IMPRS astrophysics introductory course, Lecture 7; Figure 10.4: Illustris Collaboration/Illustris Simulation; Figure 10.5: Springel, V. Max Planck Institute for Astrophysics; Figure 10.6: Vikhlinin, A. *et al.* (2009) ‘Chandra cluster cosmology project III: cosmological parameter constraints’, *The Astrophysical Journal* 692(2), IOP Publishing; Figure 11.2: The THESAN Collaboration/HTML5 UP, https://www.thesan-project.com/images/media/thesan_lightcone.png, this work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.4: Klessen, R. S. *et al* (2012) ‘On the formation of very metal poor stars: the case of SDSS J1029151+172927’, *Monthly Notices of the Royal Astronomical Society*, 421(4), pp. 3217–3221, Oxford University Press; Figure 11.5a, b and c: Marcelo Alvarez (CITA), Ralf Kaehler (Stanford), Tom Abel (Stanford); Figure 11.8: Cmglee, https://commons.wikimedia.org/wiki/File:Nucleosynthesis_periodic_table.svg, this file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0) license, <https://creativecommons.org/licenses/by-sa/3.0/deed.en>; Figure 11.9a: NASA/CXC/SAO; Figure 11.9b: NASA, ESA and the Hubble Heritage Team (STScI/AURA), Gallagher, J. (University of Wisconsin), Mountain, M. (STScI) and Puxley, P. (NSF), <https://esahubble.org/images/heic0604a/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.9c: NASA Goddard Space Flight Center; Figure 11.9d: NASA/ESA/STScI/AURA/The Hubble Heritage; Figure 11.10a: NRAO/AUI/NSF/

Univ. Hertfordshire/Hardcastle, M. <https://public.nrao.edu/gallery/centaurus-a-in-radio/>, licensed for use under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, <https://creativecommons.org/licenses/by/3.0/>; Figure 11.10b: ESO, <https://www.eso.org/public/images/eso1221a/>, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.10c: ESO/WFI (optical), MPIfR/ESO/APEX/Weiss, A. *et al.* (submillimetre), NASA/CXC/CfA/Kraft, R. *et al.* (X-ray), <https://www.eso.org/public/images/eso0903a/>, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Madau, P. and Dickinson, M. (2014) ‘Cosmic star-formation history’ *Annual Review of Astronomy and Astrophysics*, 52, pp. 415–486. Available at <https://doi.org/10.1146/annurev-astro-081811-125615>; Figure 11.13: Bullock, J. S. and Boylan-Kolchin, M. (2017) ‘Small-scale challenges to the Λ CDM Paradigm’, *Annual Review of Astronomy and Astrophysics*, 55, pp. 343—387. Available at <https://doi.org/10.1146/annurev-astro-091916-055313>; Figure 11.14: Sánchez, A. G. *et al.* (2012) ‘The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological implications of the large-scale two-point correlation function’, *Monthly Notices of the Royal Astronomical Society*, 425(1), pp. 415–437. Available at <https://doi.org/10.1111/j.1365-2966.2012.21502.x>; Figure 11.15: Eisenstein, D. J. *et al.* (2006) ‘On the robustness of the acoustic scale in the low-redshift clustering of matter’, *The Astrophysics Journal*, 664, pp. 660–674, IOP Publishing, American Astronomical Society, Institute of Physics, University of Chicago Press.

Software: ‘Python’ and the Python logos are trademarks or registered trademarks of the Python Software Foundation, used by The Open University with permission from the Foundation.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Book production contributors

Academic authors

Judith Croston (Chair), Hugh Dickinson, Iain McDonald and Sheona Urquhart.

The authors would like to thank Stephen Serjeant, Bonny Barkus, Kate Gibson and Mark Jones for useful feedback and discussions.

External assessor

Stephen Wilkins, University of Sussex.

Curriculum team

Jessica Bartlett and Shelah Surveye.

Production team

Senior project manager

Jeni Aldridge.

Editors

Jonathan Martyn, Peter Twomey, Yon-Hee Kim and Lil Davies.
Mark Radford (Pepperhouse Editorial) and Jonathan Darch.

Graphics

Sha'ni Hirschy.

OU Library

James Salter.

Index

Note: **bold** page numbers indicate where terms are defined.

Abell 1689 99
 absorption line 89
 abundance 88
 abundance ratio 90
 acoustic oscillations **6**, 33
 acoustic peak **35**, 41, 44, 46, 50
 acoustic scale **42**, 44
 active galactic nucleus (AGN) **141**, 146
 AGN *see* active galactic nucleus
 AGN feedback **146**
 Aitoff projection **19**
 Andromeda 110
 angular power spectrum **23**, 25, 30
 annihilation **62**, 70, 72
 antibaryon 62
 antilepton 62
 antiquark 62
 archaeology
 galactic 147
 associated Legendre polynomial **19**
 astration **91**, 94

BAOs *see* baryon acoustic oscillations
 baryogenesis 70
 baryon acoustic oscillations (BAOs) 54, **155**
 baryon number **62**
 baryon-to-photon ratio **71**, 84
 baryonic matter 31
 BBN *see* big bang nucleosynthesis
 beta decay **63**
 BICEP Array 6
 BICEP3 6
 big bang nucleosynthesis (BBN) **65**, 74, 91
 binary star 145
 black hole 62, 141, 145
 supermassive 146
 Boltzmann factor **73**
 bound halo **120**
 bremsstrahlung radiation **13**
 Bullet cluster 103

Centaurus A 146
 chemical enrichment **90**, 141
 clustering of galaxies 154

CMB *see* cosmic microwave background
 CMB anisotropies **6**
 CMB rest frame 10
 CMB solar dipole **9**
COBE *see* *Cosmic Background Explorer*
 cold dark matter **107**
 Coma cluster 99, 102
 Compton scattering **65**
 cooling rate **133**
 cooling timescale **135**
 correlation function
 two-point 154
Cosmic Background Explorer (COBE) 7, 8
 cosmic cycle of matter 141
 cosmic microwave background (CMB) 3
 angular power spectrum 23, 25, 30
 anisotropies 6
 energy density 40
 rest frame 10
 solar dipole 9
 temperature 6, 71
 temperature fluctuation 15
 temperature fluctuation amplitude 18
 cosmic neutrino background 70
 cosmic variance **24**
 cosmic web 122
 cosmological constant 31, 50
 curvature of Universe 44

dark ages 131, 132
 dark matter 31, 58, 99
 cold 107
 evidence 103, 106
 hot 107
 dark-matter halo 120, 130
 decouple **61**, 72, 87
 degenerate parameter **53**
 density parameter 31, 33, 42, 44, 46, 49, 50
 deuterium 65, 77
 deuterium abundance 91, 94
 deuterium bottleneck **76**, 79
 deuterium formation 74
 deuteron 65, 74
 deuteron formation 65

- dissociation 63
Doppler peak *see* acoustic peak
Doppler shift 10, 40, 101
electric dipole moment **13**
electromagnetic interaction 59
electron capture **64**
electroweak transition 69
electroweak unification **68**
elemental abundance **88**
emission line 89
enrichment 90, 141
epoch of last scattering 4, 31
epoch of reionisation **139**
estimators **23**
expansion of Universe 38, 58, 115, 156
faint-end slope **152**
filament **121**
forbidden lines **132**
Fourier expansion **18**
free-fall timescale **114**
free-free emission 13
freeze out **61**
galactic archaeology **147**
Galactic foreground emission 13
galactic outflow **143**
galaxy
 formation 140
 minimum mass 145
galaxy cluster **99**
 Abell 1689 99
 Bullet 103
 Coma 99, 102
 formation 154
 weight 99
galaxy merger 103, **141**
giant molecular clouds 13
grand unification 68
grand unified theory (GUT) **68**
gravitational lensing 103
gravitational potential well 31, 35
gravitational waves 31
grid simulation **123**
GUT *see* grand unified theory
H₂ formation heating **134**
hadron era 70
hadron formation 63
halo mass function **125**, 151, 154
helium abundance 78, 94, 95
helium formation 77
helium-3 77
helium-4 77, 80
HII regions **96**
hot dark matter 107
Hubble constant 46
Hubble diagram 110
Hubble flow 110
Hubble–Lemaître relationship 110
Hubble parameter 61, 87, 156
IGM *see* intergalactic medium
Illustris simulation 124
inertial frame 9
inflation **68**
interacting binary star **145**
interaction cross-section **59**
interaction energy **60**
intergalactic medium (IGM) **138**
interstellar dust 13, 134
interstellar medium (ISM) **13**, 91, 136, 141
inverse Compton scattering **65**
ionisation 67
ionisation fraction 84
ISM *see* interstellar medium
jansky (Jy) 7
Jeans criterion **111**, 112
Jeans length **111**
Jeans mass **111**, 135
jet 141, **146**
Jupiter 94
Keplerian rotation **105**
Large Magellanic Cloud (LMC) 148
last scattering *see* epoch of last scattering,
 surface of last scattering
Legendre polynomial 19
 evaluation 21
lepton number **62**
lithium 80
lithium abundance 97
LMC *see* Large Magellanic Cloud
Local Group **10**, 110, 114
Lorentz transformations 10

- luminosity function **150**, 154
 Lyman limit **136**
 major merger 141
 mass deficit **66**
 mass fraction **78**, 90
 mass segregation **145**
 merger
 galaxy 103, 141
 stellar 65
 Messier 31 110
 Messier 82 143
 metallicity **90**, 96
 metals 90, 133
 Milky Way 10, 17, 92–94, 110, 114, 148
 total mass 106
 Millennium Simulation 125
 minor merger 141
 moving-mesh simulation **123**
 multipole number 19
N-body simulation **123**
 nested sampling 123
 neutrino decoupling 70
 neutron star 145
 neutron-to-proton ratio 72, 78
 Newton's shell theorems 113
 NGC 1560 105
 NGC 5128 146
 nucleosynthesis *see* big bang nucleosynthesis
 number counts **149**
 numerical simulation 122, *see also* simulation
 overdensity 10, 31, 35, 39, 112, 114, 118, 130,
 156
 pair production **61**, 62
 parameter degeneracies **53**
 photodisintegration **66**, 74
 photoionisation 67
 photon–baryon fluid **33**, 33–36, 40, 155
Planck 7–9, 16, 91, 92
 Planck time **68**
 Population I **136**
 Population II **136**, 142
 Population III **136**, 142
 positron capture **64**
 primordial nucleosynthesis **74**, *see* big bang
 nucleosynthesis
 protogalaxy **140**
 quasar 91
 quenching **146**
 radiation driving **39**, 49
 radio lobe 141, **146**
 reaction rate **59**
 recombination 67, 83–88, 132, 137–138
 recombination coefficient **137**
 red and dead 145
 red giant 142
 reionisation **136**, 139
 relative abundance **90**
 RMS *see* root mean square
 root mean square (RMS) **14**
 rotation curve 104
 Sachs–Wolfe effect **39**
 Sachs–Wolfe plateau **32**
 Sagittarius 148
 Saha equation 67, 84
 Saturn 94
 scale factor 68, 115, 161
 Sculptor 148
 SDSS *see* Sloan Digital Sky Survey
 shell theorems 113
 Silk damping **40**
 simulation 107, 122–126, 139
 Illustris 124
 Millennium Simulation 125
 types 122
 Sloan Digital Sky Survey (SDSS) 89
 SMBH *see* supermassive black hole
 smoothed-particle hydrodynamic (SPH)
 simulation **123**
 sound horizon **42**, 155
 South Pole Telescope 6
 spectral line 90
 spectroscopy **89**
 spherical harmonics **18**, 18–22
 standard deviation **93**
 standard error on the mean **93**
 star formation efficiency **153**
 Stefan–Boltzmann law 40
 stellar feedback **143**
 stellar mass **150**
 stellar mass function **150**, 150–154

Index

stellar merger 65
stellar remnant **145**
Strömgren sphere **137**
 radius 137
strong interaction 59
structure formation 107, 117, 149
super-horizon perturbations 32
supermassive black hole (SMBH) **146**
supernova 136, 145
 Type Ia 54
surface of last scattering **4**, 10
synchrotron radiation **13**, 146

Thomson cross-section 87
Thomson scattering 3, 33, 64, 86
tidal stream **141**
tracer 106, 136, 147, 151
tritium 77
two-point correlation function **154**
Type Ia supernova 54

unification 68

velocity dispersion **101**
virial theorem **100**, 132
virialisation 120
volume integral 100

weak interaction 59, 63
white dwarf 145
Wilkinson Microwave Anisotropy Probe
 (*WMAP*) 7, 8, 53
WMAP *see* *Wilkinson Microwave Anisotropy Probe*

Zeldovich pancake **121**

S385 Errata: Cosmology Part 2

Chapter 6

- p17. There are several references on this page to Equation 5.15 that should instead refer to Equation 5.14 from *Cosmology Part 1* on page 126, which defines the angular diameter distance (d_A) in terms of the present-day proper distance ($d_p(t_0)$) and the redshift (z).
- p21. In the boxed section about Legendre polynomials, the expressions for calculating the number of nodes and extrema given values of m and ℓ are only correct if $|m| > 0$. For the $m = 0$ case Legendre polynomials have ℓ nodes and $\ell - 1$ extrema.
- p22. In the unnumbered equation giving an expression for $a_{\ell m}$ at the top of the page, the limits for the outer integral should be 0 and 2π and the limits for the inner integral should be 0 and π . The correct expression is therefore:

$$a_{\ell m} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) Y_{\ell m}^*(\theta, \phi) d\Omega$$

where $d\Omega = \sin \theta d\theta d\phi$.

- p32. The reference to Example 6.2 in the sentence immediately following Example 7.1 should instead be a reference to Example 6.3.
- p36. The sentence immediately following the blue box incorrectly states “The physical sizes of the acoustic oscillations were directly proportional to their oscillation frequencies.” The correct statement is that “The physical sizes of the acoustic oscillations were *inversely* proportional to their oscillation frequencies.”
- p166. In the solution to Exercise 11.3, when calculating R_{\max} in scenario (ii) there are two mistakes. First, in the expression for the conversion between cm^3 and m^3 the factor of 10^{-6} on the right-hand of the equals sign should be 10^{-5} . Then, in the following expression the numerator of the fraction inside the square brackets should be $3 \times 10^{55} \text{s}^{-1}$. Rectifying these two issues means that the final answer for part (a)(ii) becomes 13 Mpc. Propagating this new answer to part (c)(ii) gives $R(t) \approx 0.65 \text{ Mpc}$.

Solutions to exercises

Solution to Exercise 1.1

As set out in Equation 1.4, the surface brightness we observe for an object of a particular angular size depends on its flux. So we must first calculate the Sun's flux from its luminosity using Equation 1.3:

$$\begin{aligned} F_{\odot} &= L_{\odot}/4\pi D^2 \\ &= (3.8 \times 10^{26} \text{ W})/(4\pi(1.5 \times 10^{11} \text{ m})^2) \end{aligned}$$

which works out to be approximately 1340 W m^{-2} . We can now use Equation 1.4 to determine Σ_{\odot} :

$$\begin{aligned} \Sigma_{\odot} &= \frac{F_{\odot}}{\theta_{\odot}^2} \\ &= (1340 \text{ W m}^{-2})/(0.5 \text{ deg})^2 \\ &= 5400 \text{ W m}^{-2} \text{ deg}^{-2} \end{aligned}$$

Comparing this surface brightness with the value provided for the night sky, $\sim 2 \times 10^{-13} \text{ W m}^{-2} \text{ deg}^{-2}$, we conclude that the Sun's surface brightness is $\sim 3 \times 10^{16}$ times brighter than the night sky.

Solution to Exercise 1.2

The mean free path for a photon in the halo of the Milky Way at the present time is given by

$$\begin{aligned} \lambda &= \frac{1}{(100 \text{ m}^{-3})(6.652 \times 10^{-29} \text{ m}^2)} \\ &= 1.50 \times 10^{26} \text{ m} \end{aligned}$$

The mean free path for a photon at the given epoch in the early Universe is

$$\begin{aligned} \lambda &= \frac{1}{(5.0 \times 10^9 \text{ m}^{-3})(6.652 \times 10^{-29} \text{ m}^2)} \\ &= 3.01 \times 10^{18} \text{ m} \end{aligned}$$

Converting both values into units of kpc gives $4.9 \times 10^6 \text{ kpc}$ and 0.097 kpc , respectively.

Hence the mean free path in the halo of the Milky Way is much larger than the size of a typical galaxy, whereas in early Universe conditions the mean free path was much smaller than a typical galaxy.

Solution to Exercise 1.3

We start from the simplified form of the ionisation fraction equation:

$$\frac{1-X}{X} = n_p C(T)$$

which rearranges to:

$$1 = n_p C(T) X + X = X(n_p C(T) + 1)$$

and so

$$X = \frac{1}{n_p C(T) + 1}$$

We can now substitute in the values provided for each case. First, evaluating $C(T)$:

$$C(T) = \left(\frac{m_e k_B T}{2\pi \hbar^2} \right)^{-3/2} \exp \left(\frac{Q}{k_B T} \right)$$

and taking care to use $\hbar = h/(2\pi)$, gives values of $4.8 \times 10^{-31} \text{ m}^3$, $2.2 \times 10^{-10} \text{ m}^3$ and $1.8 \times 10^{-4} \text{ m}^3$ for scenarios (a), (b) and (c), respectively.

Using our values of $C(T)$ and the given values of n_p we find that $X = 1.0$, 0.48 and 1.1×10^{-6} (i.e. close to zero), respectively, for situations (a), (b) and (c). Note that (as will be the case throughout the module) your calculations may differ very slightly depending on the precision of constants and any intermediate rounding.

Solution to Exercise 1.4

We can use the provided proportionality to compare ratios of quantities at different times:

$$\frac{T_{\text{CMB}}}{T_{\text{present}}} = \frac{1+z_{\text{CMB}}}{1+z_{\text{present}}}$$

We know the temperature of the CMB at the time of emission, $T_{\text{CMB}} \approx 3000 \text{ K}$, and its temperature as observed at the Earth in the present day, $T_{\text{present}} = 2.7 \text{ K}$. The current redshift is zero. We can therefore rearrange the equation to obtain

$$z_{\text{CMB}} = \frac{T_{\text{CMB}}}{T_{\text{present}}} - 1$$

and substituting in the provided values gives $z_{\text{CMB}} = 1110$.

Solution to Exercise 2.1

(a) If $V = 20 \text{ km h}^{-1} = 5.6 \text{ m s}^{-1}$ then, using Equation 2.5, $\gamma = 1.0$.

(b) For $V = 0.9c$ you should have determined that $\gamma = 2.3$.

If you tried the optional Python task then you should have produced a plot similar to Figure S1.

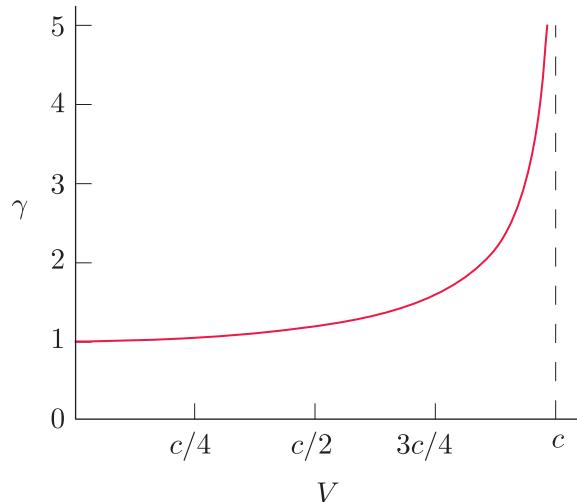


Figure S1 The dependence of the Lorentz factor, γ , on speed V .

Solution to Exercise 2.2

Equations 2.6 and 2.7 state

$$\begin{aligned}\Delta x' &= \gamma(\Delta x - V\Delta t) \\ \Delta t' &= \gamma(\Delta t - V\Delta x/c^2)\end{aligned}$$

To express these equations in terms of coordinates in the S' frame, first rearrange the second equation for Δt :

$$\Delta t = \frac{\Delta t'}{\gamma} + \frac{V\Delta x}{c^2}$$

This expression includes Δx , which is not a measurement from S' , so we need to rearrange the other interval transform so that we can eliminate Δx :

$$\Delta x = \frac{\Delta x'}{\gamma} + V\Delta t$$

Substituting this for Δx in the previous expression:

$$\Delta t = \frac{\Delta t'}{\gamma} + \frac{V(\Delta x'/\gamma + V\Delta t)}{c^2}$$

Collecting the Δt terms together gives:

$$\Delta t(1 - V^2/c^2) = \frac{1}{\gamma}(\Delta t' + V\Delta x'/c^2)$$

Recognising that $1 - V^2/c^2 = 1/\gamma^2$ and rearranging gives:

$$\Delta t = \gamma(\Delta t' + V\Delta x'/c^2)$$

which is an expression for Δt only in terms of S' coordinates.

A similar process leads to an expression for Δx in terms of only $\Delta x'$ and $\Delta t'$: $\Delta x = \gamma(\Delta x' + V\Delta t')$

Compared to the starting expressions for $\Delta t'$ and $\Delta x'$, you will see that only the sign of V has changed. This makes sense, because there is no reason we could not have labelled the two frames the other way round initially, in which case the direction of the velocity would have changed sign.

Solution to Exercise 2.3

Start with an expression for ds' in the situation where the y - and z -separations can be neglected:

$$ds'^2 = c^2 dt'^2 - dx'^2$$

The aim is now to show this is the same as ds^2 .

We can take dt' and dx' to be equivalent to the space and time intervals referred to as $\Delta t'$ and $\Delta x'$ in Equations 2.6 and 2.7, and so substitute in:

$$dx' = \gamma(dx - V dt)$$

and

$$c dt' = \gamma(c dt - V dx/c)$$

where we have multiplied the differential form of Equation 2.7 by c , so that

$$ds'^2 = [\gamma(c dt - V dx/c)]^2 - [\gamma(dx - V dt)]^2$$

which simplifies to

$$ds'^2 = \gamma^2 c^2 dt^2 (1 - V^2/c^2) - \gamma^2 dx^2 (1 - V^2/c^2)$$

Recalling the definition of $\gamma = 1/\sqrt{1 - V^2/c^2}$, this becomes

$$ds'^2 = c^2 dt^2 - dx^2 = ds^2$$

as required.

Solution to Exercise 2.4

The distance along a path on the surface of a sphere can be calculated by integrating the line element dl along a parametrisation of the path. For a line of constant latitude, this is very similar to the calculation in Example 2.6 because lines of latitude have a constant angle θ if the coordinate system is defined as in Figure 2.11, so that the angle ϕ corresponds to longitude, and θ corresponds to 90° minus the latitude value.

The distance, L , is therefore given by

$$L = \int_{\phi_1}^{\phi_2} R \sin \theta \, d\phi$$

where ϕ_1 and ϕ_2 are the start and end values of longitude, and $R = 6370$ km is the (average) radius of the Earth.

We need to work with angles in radians (rad), and so $\theta = 90^\circ - 51^\circ = 39^\circ$, which corresponds to 0.681 rad. The actual longitude angles were not provided, but the location of the zero value doesn't affect the calculation, so we can define $\phi_1 = 0^\circ$ and $\phi_2 = 118^\circ = 2.06$ rad.

Solving the integral as in Example 2.6 gives

$$\begin{aligned} L &= R \sin \theta [\phi]_{\phi_1}^{\phi_2} = R \sin \theta \phi_2 \\ &= (6370 \text{ km})(\sin(0.681))(2.06) = 8260 \text{ km} \end{aligned}$$

So the distance along the route at fixed latitude is 8260 km, which is around 13 per cent longer than the great-circle distance of ~ 7300 km.

Solution to Exercise 2.5

(a) We can use Equation 2.17 to calculate the radius corresponding to the provided value of k_x :

$$R = \frac{1}{k_x} = \frac{1}{0.2 \text{ cm}^{-1}} = 5 \text{ cm}$$

Because the curvature, and hence R , are the same at all locations, the curve must be a circle of radius 5 cm.

(b) A straight line must have constant curvature. But however large a circle we draw tangential to the line, an even larger circle will approximate the straight line better. The curvature of a straight line must therefore be smaller than $1/R$ for all R ,

i.e. $k_x = 0$. Mathematically,

$$k_x = \lim_{R \rightarrow \infty} \frac{1}{R} = 0$$

Solution to Exercise 3.1

The second half of the expression, ‘matter tells space how to curve’, is a rough description of Einstein's field equations: the energy–momentum tensor is the source of curvature and determines the components of the curvature tensor.

The first half of the expression, ‘space tells matter how to move’, refers to what a metric theory of gravity means: the metric (‘space’ in the colloquial expression) determines the geodesics along which, in the absence of external forces, test particles (‘matter’) must move.

Solution to Exercise 3.2

Using Equation 3.3, we find:

Table S1 Schwarzschild and physical radii values for three objects.

Object	R_S/m	R/m
Earth	8.9×10^{-3}	6.4×10^6
Sun	3.0×10^3	7.0×10^8
neutron star	5.9×10^3	1.5×10^4

The table therefore shows that the Schwarzschild radii of both the Earth and the Sun are much smaller than their actual size. For a neutron star – the densest type of star known – the Schwarzschild radius is still less than the star's actual radius, but only by a small factor.

Solution to Exercise 3.3

(a) A black hole of mass $3 M_\odot$ has a Schwarzschild radius of 9 km. Substituting this into Equation 3.6 gives a time to reach the singularity of 2×10^{-5} s.

(b) A 10^9 -solar-mass black hole has a Schwarzschild radius of $\sim 3 \times 10^{12}$ m, which results in a travel time of ~ 7000 s.

Solution to Exercise 3.4

Taking the scale factor at the present time, $a(t_0)$, to be 1, and $H(t) = H_0$, Equation 3.11 rearranges to

$$\frac{da}{dt} = H_0$$

and so

$$\int_{a(t)}^{a(t_0)} da = H_0 \int_t^{t_0} dt$$

so that

$$a(t_0) - a(t) = 1 - a(t) = H_0(t_0 - t)$$

So for part (a) we know that $a(t) = 1 - H_0(1000 \text{ y})$. To evaluate this, the units of H_0 and t need to be converted to match each other.

H_0 can be expressed in units of s^{-1} as follows:

$$H_0 = \frac{68\,000 \text{ m s}^{-1} \text{ Mpc}^{-1}}{3.086 \times 10^{22} \text{ m Mpc}^{-1}} = 2.2 \times 10^{-18} \text{ s}^{-1}$$

Likewise, 1000 years is $3.2 \times 10^{10} \text{ s}$, while 100 million years is $3.2 \times 10^{15} \text{ s}$.

Hence for part (a), $a(t) = 1 - 7.0 \times 10^{-8}$, which is very close to 1, and for part (b), $a(t) = 1 - 0.0070 = 0.993 \approx 0.99$ to the accuracy of this calculation.

You can therefore conclude that, with the assumption that H_0 is constant, the scale factor changed by a tiny factor over the past millennium, and by a little less than 1% ($\sim 0.7\%$) over the timescale of 100 million years.

Solution to Exercise 3.5

Using the relationship between frequency and Δt , Equation 3.14 becomes

$$\frac{1}{a(t_{\text{em}})\nu_{\text{em}}} = \frac{1}{a(t_{\text{obs}})\nu_{\text{obs}}}$$

which rearranges to give a relation for the ratio between scale factors at the time light was emitted and observed:

$$\frac{a(t_{\text{obs}})}{a(t_{\text{em}})} = \frac{\nu_{\text{em}}}{\nu_{\text{obs}}}$$

Recognising that the right-hand ratio is related to redshift via Equation 3.13, we can obtain the

required relationship between a and z :

$$\frac{a(t_{\text{obs}})}{a(t_{\text{em}})} = 1 + z \quad (\text{S1})$$

Solution to Exercise 3.6

The ratios of the scale factor at the time the light from each galaxy was emitted to its value at the present time are calculated via Equation 3.15 to be $a_{\text{em}}/a_0 = 1/(1+10.1) = 0.09$ and $a_{\text{em}}/a_0 = 1/(1+3.6) = 0.22$.

Therefore, the distance between any two locations was around 9% of its current distance at $z = 10.1$ and around 22% of its current distance at $z = 3.6$.

Solution to Exercise 4.1

For a group of n non-relativistic particles, each with energy E , contained within an arbitrary volume V of the homogeneous fluid, we can write the energy density as

$$\epsilon = \frac{nE}{V} \quad (\text{S2})$$

Now, if $v \ll c$ then $p \approx m_0 v$ (where v is the speed of the particles), and we can write the particles' energy as

$$E = \sqrt{p^2 c^2 + m_0^2 c^4} \approx m_0 c^2 \sqrt{1 + \frac{v^2}{c^2}}$$

Taking the first two terms in the Taylor expansion for $\sqrt{1+x}$ with $x = v^2/c^2$, and remembering that $v \ll c$, yields

$$E \approx m_0 c^2 + \frac{1}{2} m_0 v^2 \approx m_0 c^2$$

Substituting for E in Equation S2 we can rewrite the energy density of our particles as

$$\epsilon = \frac{nE}{V} \approx \frac{nm_0 c^2}{V} \approx \rho c^2$$

Solution to Exercise 4.2

(a) We are told that the WHIM behaves like an ideal gas, so we can use the ideal gas law to compute an approximate value for w . Comparing the form of Equation 4.13 with that of Equation 4.11, substituting μ for the proton mass m_p and using the proton temperature T_p , we can

compute w using

$$\begin{aligned} w &= \frac{k_B T}{\mu c^2} = \frac{k_B T_p}{m_p c^2} \\ &= \frac{1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^7 \text{ K}}{1.673 \times 10^{-27} \text{ kg} \times (2.998 \times 10^8 \text{ m s}^{-1})^2} \\ &= 9.184 \times 10^{-7} \end{aligned}$$

So $w < 10^{-6}$, which agrees with the assertion made in the question.

(b) We have shown that $w < 10^{-6}$ for protons in the WHIM, but we ultimately want to determine whether their velocities are non-relativistic.

If we assume that the gas *is* non-relativistic, then the average kinetic energy of a particle is $(1/2)\langle \mu v_p^2 \rangle = (1/2)\mu \langle v_p^2 \rangle$. We can then write:

$$3k_B T_p = \mu \langle v_p^2 \rangle = m_p \langle v_p^2 \rangle \quad (\text{S3})$$

where $\langle v_p^2 \rangle$ is the mean squared velocity of the protons in the WHIM. If we find that $\langle v^2 \rangle \ll c^2$ then our initial assumption will have been justified. Rearranging Equation S3 we find:

$$\begin{aligned} \langle v_p^2 \rangle &= \frac{3k_B T_p}{m_p} \\ &= \frac{3 \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^7 \text{ K}}{1.673 \times 10^{-27} \text{ kg}} \\ &= 2.476 \times 10^{11} \text{ J kg}^{-1} \\ &= 2.476 \times 10^{11} \text{ m}^2 \text{ s}^{-2} \\ &\approx 3 \times 10^{-6} c^2 \end{aligned}$$

(c) The previous step shows that $\langle v^2 \rangle$ is indeed much smaller than c^2 in this example. This result justifies the assumption that even the superheated protons in the WHIM are non-relativistic.

Solution to Exercise 4.3

When $k = 0$, the Friedmann equation becomes a separable first-order differential equation. To solve for $a(t)$, we start by rearranging Equation 4.16 to separate terms related to a and t :

$$a^{1/2} da = \left(\frac{8\pi G}{3} \rho_{m,0} \right)^{1/2} dt$$

The next step is to integrate both sides, and after collecting constants of integration we obtain

$$\frac{2}{3} a^{3/2} = \left(\frac{8\pi G}{3} \rho_{m,0} \right)^{1/2} t + \cancel{\text{constant}}^0$$

In case it is unfamiliar, the notation \cancel{X}^0 means ‘ X cancels to zero’. In this instance, we can set the integration constant to zero by assuming that the universe in question has zero size at the big bang when $t = 0$, and therefore that $a(0) = 0$. The expression can then be rearranged to obtain

$$a = (6\pi G \rho_{m,0})^{1/3} t^{2/3}$$

Finally, we use the boundary condition that $a(t_0) = 1$ to show that

$$a(t) = \left(\frac{t}{t_0} \right)^{2/3}$$

when $t_0 = (6\pi G \rho_{m,0})^{-1/2}$, as required.

Solution to Exercise 4.4

We can use exactly the same approach as was used for a matter-only universe. To solve for $a(t)$, we start by rearranging Equation 4.21 to separate terms related to a and t :

$$a da = \left(\frac{8\pi G}{3} \rho_{r,0} \right)^{1/2} dt$$

The next step is to integrate both sides, and after collecting constants of integration we obtain

$$\frac{a^2}{2} = \left(\frac{8\pi G}{3} \rho_{r,0} \right)^{1/2} t + \cancel{\text{constant}}^0$$

As in the matter-only case, we can set the integration constant equal to zero by assuming that the universe in question has zero size at the big bang when $t = 0$, and therefore that $a(0) = 0$. Then this expression can be rearranged to obtain

$$a = \left(\frac{32\pi G \rho_{r,0}}{3} \right)^{1/4} t^{1/2}$$

Finally, we use the boundary condition that $a(t_0) = 1$ to show that

$$a(t) = \left(\frac{t}{t_0} \right)^{1/2}$$

when $t_0 = (32\pi G \rho_{r,0}/3)^{-1/2}$.

Solution to Exercise 4.5

In general, the curvature parameter k can be positive, negative or zero. If a universe is static then $\dot{a} = 0$, and Equation 4.34 can be written as

$$\frac{8\pi G}{3}\rho + \frac{\Lambda c^2}{3} = \frac{kc^2}{a^2}$$

You are told to consider a universe that is not empty, so $\rho > 0$, and the question indicates that $\Lambda > 0$. Consequently, both terms on the left-hand side of this rewritten equation are non-zero and positive, so the right-hand side must also be positive. We also know that $a > 0$ except when $t = 0$ at the big bang, so the curvature k must be positive too.

Solution to Exercise 5.1

If we adopt the standard convention that $a(t_0) = 1$, then we can simplify Equation S1 from Chapter 3 and relate the redshift of photons that we detect today to the scale factor of the Universe at the time t , when they were emitted:

$$1+z = \frac{1}{a(t)} \quad (\text{S4})$$

The Friedmann equation is given in Equation 4.41 as:

$$\frac{H^2}{H_0^2} = \frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{r,0}}{a^4} + \Omega_{\Lambda,0} + \frac{\Omega_{k,0}}{a^2}$$

Substituting for a using Equation S4 then taking the square root and rearranging, we find

$$\begin{aligned} H &= H_0 [\Omega_{m,0}(1+z)^3 + \Omega_{r,0}(1+z)^4 \\ &\quad + \Omega_{\Lambda,0} + \Omega_{k,0}(1+z)^2]^{1/2} \\ &= H_0 E(z) \end{aligned} \quad (\text{S5})$$

Now, by differentiating Equation S4 and rearranging, we find:

$$da = -a(t)^2 dz$$

We also know that:

$$H = \frac{\dot{a}}{a} = \frac{1}{a(t)} \frac{da}{dt}$$

Rearranging this expression and substituting for H using Equation S5 gives:

$$dt = \frac{da}{Ha(t)} = -\frac{a(t)}{E(z)H_0} dz$$

Finally, substituting for dt in Equation 5.7 gives:

$$\begin{aligned} d_p(t_0) &= c \int_{t_{\text{em}}}^{t_0} \frac{dt}{a(t)} \\ &= -c \int_{z(t_{\text{em}})}^{z(t_0)} \frac{a(t)}{a(t)E(z)H_0} dz \\ &= -\frac{c}{H_0} \int_{z_{\text{em}}}^0 \frac{dz}{E(z)} \\ &= \frac{c}{H_0} \int_0^{z_{\text{em}}} \frac{dz}{E(z)} \end{aligned}$$

Solution to Exercise 5.2

Consider a particular object at redshift z . By rearranging and combining Equations 5.12 and 5.14, the current proper distance to the object is related to d_L and d_A by:

$$d_p(t_0) = (1+z) d_A = \frac{d_L}{1+z}$$

Now let the observed angular size and flux of the object be θ_{obs} and F_{obs} , respectively. Using the definitions of d_L and d_A , we obtain:

$$(1+z) \frac{l}{\theta_{\text{obs}}} = \frac{1}{1+z} \sqrt{\frac{L}{4\pi F_{\text{obs}}}} \quad (\text{S6})$$

We are interested in demonstrating proportionality, so let's approximate the observed appearance of our object as a disc with diameter θ_{obs} . In that case, the solid angle subtended by the disc is proportional to θ_{obs}^2 . Rearranging Equation S6, we find the required proportionality relationship:

$$\frac{F_{\text{obs}}}{\theta_{\text{obs}}^2} = \frac{L}{4\pi l^2} \frac{1}{(1+z)^4} \quad (\text{S7})$$

Solution to Exercise 5.3

We will assume that d_b is the mean Earth–Sun distance, i.e. $d_b \approx 1 \text{ AU}$. Rearranging Equation 5.19 to isolate θ_{ϖ} we find:

$$\theta_{\varpi} = \frac{d_b}{d_p} \approx \frac{1 \text{ AU}}{163\,000 \text{ ly}}$$

Performing the calculation and converting units appropriately we find that:

$$\begin{aligned} \theta_{\varpi} &\approx 0.97 \times 10^{-10} \text{ radians} \\ &\approx 0.02 \text{ milliarcseconds} \\ &\approx 2 \times 10^{-5} \text{ arcseconds} \end{aligned}$$

Solution to Exercise 5.4

The value of H_0 provided has *dimensions* of inverse time (i.e. s^{-1}), but its *unit* is expressed in terms of two different length scales: Mpc and km. To eliminate the unit's dependence on distance we simply divide by the number of km in 1 Mpc, which is approximately 3.086×10^{19} .

This gives

$$\begin{aligned} H_0 &\approx \frac{67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}}{3.086 \times 10^{19} \text{ km Mpc}^{-1}} \\ &= 2.194 \times 10^{-18} \text{ s}^{-1} \end{aligned}$$

So the Hubble time is approximately

$$t_H \approx \frac{1}{2.194 \times 10^{-18} \text{ s}^{-1}} = 4.558 \times 10^{17} \text{ s}$$

Now we just need to convert this to My. There are 31 557 600 seconds in 1 year, so

$$t_H = \frac{4.558 \times 10^{17}}{31 557 600 \times 10^6} = 14 443 \text{ My}$$

or approximately 14.5 billion years. This value is on a similar scale to the current best estimate of the Universe's true age, which is around 13.7 billion years.

Solution to Exercise 5.5

By rearranging Equation 5.25 to isolate the apparent magnitude we can write:

$$\begin{aligned} m &= M + 5 \log_{10}(d_L/\text{pc}) - 5 \\ &= M + 5 \log_{10}\left(\frac{d_L/\text{pc}}{10/\text{pc}}\right) \end{aligned}$$

The units of H_0 mean that it will be more convenient if we express d_L in Mpc and so our previous expression becomes:

$$\begin{aligned} m &= M + 5 \log_{10}\left(\frac{d_L/\text{Mpc}}{10^{-5}/\text{Mpc}}\right) \\ &= M + 5 \log_{10}(d_L/\text{Mpc}) + 25 \end{aligned}$$

Now we use an algebraic trick of multiplying d_L by H_0/H_0 and rearranging:

$$\begin{aligned} m &= M + 5 \log_{10}\left(\frac{H_0}{H_0} d_L/\text{Mpc}\right) + 25 \\ &= M - 5 \log_{10} H_0 + 5 \log_{10}(H_0 d_L/\text{Mpc}) + 25 \end{aligned}$$

Solution to Exercise 5.6

We want to find the value of z for which:

$$1.1 \times cz = cz \left(1 + \frac{1 - q_0}{2} z\right)$$

Rearranging to isolate z , we find:

$$z = \frac{0.2}{1 - q_0}$$

We are told to assume that the Universe is flat and that $\Omega_{r,0}$ is negligible so we can use Equation 4.46 to compute q_0 .

$$\begin{aligned} q_0 &= \frac{\Omega_{m,0}}{2} - \Omega_{\Lambda,0} \\ &= \frac{0.3097}{2} - 0.6888 \\ &= -0.5340 \end{aligned}$$

Using this value for q_0 in Equation , we find:

$$z = \frac{0.2}{1 + 0.5340} = 0.1304$$

so the low-redshift approximation is 10% too low when $z = 0.13$.

Solution to Exercise 5.7

We know that all galaxies contain CCs, and that roughly one Type Ia supernova occurs per decade in a spherical volume with proper radius $r_p = 20 \text{ Mpc}$. We also know that if $z \ll 1$, then $d_p(t_0) \approx cz/H_0$. Using this approximation we can calculate that CCs are detectable within a volume that has proper radius:

$$\begin{aligned} r_p(z = 0.01) &\approx \frac{zc}{H_0} \approx \frac{0.01 \times 2.998 \times 10^5 \text{ km s}^{-1}}{67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}} \\ &\approx 44.3 \text{ Mpc} \end{aligned}$$

Now our expected supernova rate per decade is just the ratio of two spherical volumes

$$\frac{(44.3 \text{ Mpc})^3}{(20 \text{ Mpc})^3} \approx 11$$

We would expect to find just over one Type Ia supernova every year in a galaxy that contains detectable CCs.

References and acknowledgements

References

- Benedict, G. F. *et al.* (2007) ‘*Hubble Space Telescope* fine guidance sensor parallaxes of galactic Cepheid variable stars: period–luminosity relations’, *The Astronomical Journal*, 133(4), pp. 1810–1827. Available at <https://doi.org/10.1086/511980>.
- Betoule, M. *et al.* (2014) ‘Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples’, *Astronomy & Astrophysics*, 568, A22. Available at <https://doi.org/10.1051/0004-6361/201423413>.
- Hamuy, M. *et al.* (1993) ‘The 1990 Calán/Tololo supernova search’, *The Astronomical Journal*, 106(6), pp. 2392–2407. Available at <https://doi.org/10.1086/116811>.
- Kolb, U. (2010) *Extreme environment astrophysics*. Cambridge: Cambridge University Press, in association with The Open University.
- Lambourne, R. (2010) *Relativity, gravitation and cosmology*. Cambridge: Cambridge University Press, in association with The Open University.
- Liddle, A. (1999) *An introduction to modern cosmology*. Chichester: Wiley.
- Misner, C., Thorne, K. and Wheeler, J. (1973) *Gravitation*. San Francisco: W. H. Freeman.
- Newman, J. A. *et al.* (1999) ‘A Cepheid distance to NGC 4603 in Centaurus’, *The Astrophysical Journal*, 523(2), pp. 506–520. Available at <https://doi.org/10.1086/307764>.
- Perlmutter, S. *et al.* (1997) ‘Measurements of the cosmological parameters Ω and Λ from the first seven supernovae at $z \geq 0.35$ ’, *The Astrophysical Journal*, 483(2), pp. 565–581. Available at <https://doi.org/10.1086/304265>.
- Perlmutter, S. *et al.* (1997) ‘Measurements of Ω and Λ from 42 high-redshift supernovae’, *The Astrophysical Journal*, 517(2), pp. 565–586. Available at <https://doi.org/10.1086/307221>.
- Ryden, B. (2017) *Introduction to cosmology*. 2nd edn. New York: Cambridge University Press.
- Serjeant, S. (2010) *Observational cosmology*. Cambridge: Cambridge University Press, in association with The Open University.
- van Velzen, S. *et al.* (2021) ‘Seventeen tidal disruption events from the first half of ZTF survey observations: entering a new era of population studies’, *The Astrophysical Journal*, 908(4). Available at <https://doi.org/10.3847/1538-4357/abc258>.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: Max-Planck Institute for Physics.

Chapter images: Figure 1.1: M. Collness, Mount Stromlo Observatory; Figure 1.2: Hubble, E. (1929) ‘A relation between distance and radial velocity among extra-galactic nebulae’, *Proceedings of the National Academy of Sciences of the United States of America*, 15(3), pp. 168-173; Figure 1.3: Betoule, M. *et al.* (2014) ‘Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples’, *Cosmology and Nongalactic Astrophysics*, Cornel University, <https://arxiv.org/abs/1401.4064>, licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, <https://creativecommons.org/licenses/by/4.0/>; Figure 1.5: MissMJ, Cush, https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, licensed under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, <https://creativecommons.org/licenses/by/3.0/deed.en>; Figure 1.6: NASA/WMAP Science Team; Figure 1.8: ESA – C. Carreau, https://www.esa.int/ESA_Multimedia/Images/2013/03/Planck_history_of_Universe, licensed under a Creative Commons Attribution-ShareAlike 3.0 IGO (CC BY-SA 3.0 IGO) licence, <https://creativecommons.org/licenses/by-sa/3.0/igo/>; Figure 1.9: European Space Agency and the Planck Scientific Collaboration; Figure 3.6a: courtesy Caltech/MIT/LIGO Laboratory; Figure 3.6b: Kramer, M. *et al.* (2021), ‘Strong-field gravity tests with the double pulsar’, *Physics Review X*, 11:041050; Figure 3.7a: S. Gillessen, *et al.* (2009) ‘Monitoring stellar orbits around the massive black hole in the galactic centre’, *The Astrophysical Journal*, 692:10751109, American Astronomical Society, 2009; Figure 3.7b: EHT Collaboration, <https://www.eso.org/public/images/eso1907a/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 3.7c: V. Heesen and LOFAR surveys team; Figure 3.8: courtesy Caltech/MIT/LIGO Laboratory; Figure 3.11a: NASA/CXC/U. Michigan/J. Miller *et al.*, illustration NASA/CXC/M. Weiss; Figure 3.11b: Velzen, S. N. *et al.* (2021) ‘Seventeen tidal disruption events from the first half of ZTF survey observations: entering a new era of population studies’, *The Astrophysical Journal*, 908(1), The American Astronomical Society; Figure 5.6: ESA-D. Ducros, 2013; Figure 5.7: NASA/ESA, The Hubble Key Project Team and The High-Z Supernova Search Team, <https://esahubble.org/images/opo9919i/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 5.8: copyright unknown; Figure 5.12: Suzuki, N. (2012) ‘The Hubble Space Telescope Cluster Supernova Survey. V. Improving the dark-energy constraints above $z > 1$ and building an early-type-hosted supernova sample’, © 2012, The American Astronomical Society, all rights reserved, printed in the U.S.A.

Software: ‘Python’ and the Python logos are trademarks or registered trademarks of the Python Software Foundation, used by The Open University with permission from the Foundation.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Book production contributors

Academic authors

Judith Croston (Chair), Hugh Dickinson, Iain McDonald and Sheona Urquhart.

The authors would like to thank Stephen Serjeant, Bonny Barkus, Kate Gibson and Mark Jones for useful feedback and discussions.

External assessor

Stephen Wilkins, University of Sussex.

Curriculum team

Jessica Bartlett and Shelah Survey.

Production team

Senior project manager

Jeni Aldridge.

Editors

Jonathan Martyn, Peter Twomey, Yon-Hee Kim and Lil Davies.
Mark Radford (Pepperhouse Editorial) and Jonathan Darch.

Graphics

Sha’ni Hirsch.

OU Library

James Salter.

Index

Note: **bold** page numbers indicate where terms are defined.

- 2dF Galaxy Redshift Survey 4
- absolute magnitude **134**, 140
- acceleration equation 94, **94**
- age of Universe 11
- angular diameter distance **126**
- BAOs *see* baryon acoustic oscillations
- baryogenesis **22**
- baryon **14**
- baryon acoustic oscillations (BAOs) 151
- baryonic matter **14**
- big bang model **20**, 21
- big bounce universe 114
- big crunch 90
- binary pulsar system 65
- binary star system 136
- binding energy 18
- black hole 67, **69**
 - evidence 69
 - merger 70
- black-body spectrum **24**
- blueshift 6
- boson **12**
- brightness of sky 11
- Calán/Tololo Supernova Survey 145
- Cartesian coordinates 46
- causality **36**, 40
- causally connected **40**
- CC *see* classical Cepheid
- Chandrasekhar limit **136**
- classical Cepheid (CC) **133**
- closed universe **91**
- cluster 4
- CMB *see* cosmic microwave background
- co-moving coordinates **78**
- coasting universe 113
- co-moving radius 92
- conservation of energy 93
- contents of Universe 12
- coordinate system **29**
- coordinate time 57
- coordinates
 - Cartesian 46
 - polar 43
 - spherical 46
- cosmic microwave background (CMB) 20, **23**, 76
- cosmic time **77**
- cosmic web 4
- cosmological constant **107**
- cosmological distance ladder **142**
- cosmological principle 4, **4**, 8, 76, 121
- cosmological redshift 121
- critical density **103**
- critical redshift 128
- curvature **45**, 54
 - negative 48
 - positive 48
- curvature parameter **48**, **79**, 90
- curvature scalar 63
- dark ages **22**
- dark energy **15**
- dark matter **15**
- dark star 69
- deceleration parameter **112**
- δ Cephei 134
- density 15
 - critical 103
- density parameter **105**
- differential notation 42
- distance modulus **135**
- Doppler shift 69, 121
- Earth 60, 61, 129, 131
- Einstein tensor **63**
- Einstein's field equations **62**
- electromagnetic force 12
- electromagnetic interaction **13**
- electron **13**
- electron number density 18
- energy 14
- energy density 91
- energy-momentum tensor **63**
- equation of state **95**
 - cosmological constant 109
 - matter 96

- radiation 99
- equation of state parameter **95**
- equivalence principle **60**
- escape speed **69**
- event **29**
- event horizon **69**, 71, 73
- Event Horizon Telescope 69
- expansion of Universe 7, 12
- experimental uncertainty 147
- fermion **12**, 13
- first law of thermodynamics 92
- first light **22**
- flare 74
- flat universe **91**
- fluid equation **92**, 93
- flux **9**
- flux density **9**
- free fall 59
- Friedmann equation **87**, 89, 105–107, 109
- fundamental observers **76**
- Gaia* 131
- galaxy evolution **22**
- Galilean transformations **32**
- gas 15
- general relativity 45, 54, 91
 - Einstein’s field equations 63
 - evidence 64
- general theory of relativity 12, 106
- geodesic **48**, 55
- gravitational lensing 65
- gravitational potential energy 93
- gravitational redshift 65, **75**
- gravitational waves **65**
- gravity
 - metric theory 62
- great circle **48**
- H–R diagram *see* Hertzsprung–Russell diagram
- hadron **14**
- Hertzsprung–Russell (H–R) diagram 133
- High-z Supernova Search Team 138
- homogeneous 4, 76, 88, 95
- horizon distance **119**, 120
- HST* *see* *Hubble Space Telescope*
- Hubble constant **6**, 7, 80
 - estimation 140
- misnomer 8
- Hubble diagram 6, 7, 150
- Hubble distance **139**
- Hubble flow **76**
- Hubble parameter 8, **81**, 103
- Hubble Space Telescope (HST)* 134
- Hubble time **139**
- Hubble–Lemaître law **6**, 121, 141
- hypersurface 77
- ideal gas law **16**, 95, 96
- inertial frame **28**
- inflation **22**
- instability strip **133**
- interval **33**
- invariance
 - spatial 90
- ionisation **18**
- ionisation fraction 18, 20
- isotropic 4, 76, 88, 95
- Kerr metric **75**
- Large Magellanic Cloud (LMC) 131, 134, 138
- Laser Interferometer Gravitational-wave Observatory (LIGO) 65
- last scattering **22**
- length contraction 35
- lepton **13**
- light 17
- light cone **40**, 55
- light curve **137**
- light-like geodesic 55
- light-year 5
- LIGO *see* Laser Interferometer Gravitational-wave Observatory
- line element **42**
- LMC *see* Large Magellanic Cloud
- local inertial frames **61**
- loitering universe 113
- lookback time **22**
- Lorentz factor **32**
- Lorentz transformations **32**, 33, 37
- luminosity **9**
- luminosity distance **122**
- LVT151012 70
- magnitude system 133
- manifold **45**

- mass density 15
 matter 14, 96
 matter–antimatter annihilation 102
 matter fluid 96
 Maxwell–Boltzmann distribution **15**
 mean free path **17**
 measured redshift 121
 measurement uncertainty 147
 megaparsec (Mpc) 5
 metric **44**, 51, 55
 metric coefficient **44**
 metric tensor **44**, 63
 metric theory of gravity 62
 Michelson–Morley experiment 29
 Milky Way 69, 78
 - peculiar velocity 131
 Minkowski metric **44**, 54, 79
 Minkowski space 56
 Minkowski spacetime 54

 negative curvature **48**
 neutrino **13**, 99
 neutron **13**
 NGC 1262 78
 NGC 4526 136
 NGC 4603 135
 non-Euclidean geometry **45**
 notation 10
 nucleosynthesis **22**
 null geodesic **55**
 number density
 - electron 18
 - particle 15
 observer **29**
 Olbers’ paradox **11**
 opacity **17**
 open universe **91**
 overdensity 92

 parsec (pc) 5, 131
 particle horizon **119**
 particle number density 15
 particles form **22**
 Pauli exclusion principle **12**
 peculiar velocity **121**
 - Milky Way 131
 perfect fluid **95**

 perihelion 64
 period–luminosity relation **134**
 photon **12**, 16, 18, 99
 plasma **13**, 15
 polar coordinates 43
 positive curvature **48**
 positron **14**
 postulates of special relativity 29
 Pound–Rebka experiment **65**
 precession **64**
 pressure gradient 94
 proper distance **118**
 proper radius *see* particle horizon
 proper time **55**, 57, 71
 proton **13**
 pseudo-Riemannian 54

 quark **13**
 quasar 65, 130

 radiation 17, 99
 radiation constant 17
 radiation fluid 99
 recession 5
 recession velocity 106
 recombination **22**
 redshift **6**, 24, 25, 81, 83, 120
 - cosmological 121
 - critical 128
 - measured 121
 reference frame **28**, 30, 34, 56, 61
 - transformation 31
 relativistic 16
 relativity of simultaneity **42**
 Ricci curvature 63
 Riemann curvature tensor **53**
 Riemannian geometry 53
 Robertson–Walker metric **79**, 90

 Saha equation **18**, 19
 scale factor **78**, 80, 83, 98, 99, 118
 Schwarzschild metric **67**, 68
 Schwarzschild radius **67**, 68
 SCP *see* Supernova Cosmology Project
 SDSS *see* Sloan Digital Sky Survey
 Shapiro delay **65**
 simultaneity 42
 singularity **71**, 76

- Sloan Digital Sky Survey (SDSS) 7
SN 1987A 138
SN 1994D 136
SN UDS10Wil 138
SNLS *see* Supernova Legacy Survey
space-like geodesic 55
spacetime 8, 12, 36, 42, 45, 62, 66, 76, 94
spacetime diagram 36, 38–41
spacetime separation 43
spatial invariance 90
special relativity 27
 applications 36
 consequences 33
 postulates 29
 terminology 29
spectral shift 6
speed of light 29
spherical coordinates 46
standard candle 122
 Type Ia supernova 136
Standard Model 12
standard rod 126
standard ruler *see* standard rod
standard yardstick *see* standard rod
standardisable candle 137
stellar parallax 130, 131, 134
stress energy 94
strong interaction 13
structure formation 22
sub-relativistic 15
Sun 11, 64, 129, 131
 notation 10
supercluster 4
supernova 7
 standard candle 136
 standardisable candle 137
 Type Ia 136, 138, 140
Supernova Cosmology Project (SCP) 138, 145, 150
Supernova Legacy Survey (SNLS) 7
surface brightness 9, 11, 128

TDE *see* tidal disruption event
tensor 44
thermal equilibrium 15, 24
Thomson cross-section 18
Thomson scattering 17
tidal disruption event (TDE) 74
time dilation 35, 75
time-like geodesic 55, 56
Type Ia supernova 136, 138, 140
 number of 138
 standard candle 136
 standardisable candle 137

uncertainty
 measurement 147
universality of free fall 59
Universe
 age 11
 contents 12
 expansion 7, 12
 model 95
universe
 closed 91
 flat 91
 open 91

Virgo interferometer 65

warm–hot intergalactic medium 96
WD *see* white dwarf
weak interaction 13
Weyl’s postulate 76
WHIM *see* warm–hot intergalactic medium
white dwarf (WD) 136, 136
world line 39, 55, 56

Solutions to exercises

Solution to Exercise 6.1

Using the fact that the real Universe is spatially flat, the proper distance between two points at time t is directly proportional to the scale factor $a(t) = (1+z)a(t_0) = 1+z$, so a region with proper size $l = 19 \text{ kpc}$ at $z_{\text{ls}} \approx 1090$ has proper size $\Delta l(1+z_{\text{ls}}) \approx 21 \text{ Mpc}$ today.

Solution to Exercise 7.1

Start by using Equation 3.15 from Chapter 3 to calculate the ratio of the scale factors corresponding to z_{ls} and z_{mr} :

$$\frac{a_{\text{ls}}}{a_{\text{mr}}} = \frac{1+z_{\text{mr}}}{1+z_{\text{ls}}} = \frac{5731}{1091} \approx 5.25 \quad (\text{S8})$$

Now we can use Equation 7.2 to calculate the ratio of ages of the model universe when $a = a_{\text{mr}}$ and when $a = a_{\text{ls}}$. Note that the coefficient *outside* the square brackets in Equation 7.2 will cancel in the ratio.

When $a = a_{\text{mr}}$ the term *inside* the square brackets reduces to $1 - \sqrt{2}/2$, and we can use the result from Equation S8 to write

$$\begin{aligned} \frac{t_{\text{mr}}}{t_{\text{ls}}} &= \frac{\left(1 - \frac{\sqrt{2}}{2}\right)}{1 - \left(1 - \frac{a_{\text{ls}}}{2a_{\text{mr}}}\right) \left(1 + \frac{a_{\text{ls}}}{a_{\text{mr}}}\right)^{1/2}} \\ &= \frac{\left(1 - \frac{\sqrt{2}}{2}\right)}{1 - \left(1 - \frac{5.25}{2}\right) (1 + 5.25)^{1/2}} \\ &= 0.0579 \end{aligned}$$

We have therefore shown that this universe had only been radiation-dominated for $\sim 5.8\%$ of its history by the epoch of last scattering. For the subsequent $\sim 94\%$ of its history, up until the time when the CMB photons were released, energy density had been dominated by matter, of which dark matter formed the dominant component. You may have reached the same conclusion by

separately calculating $t(a_{\text{ls}})$ and $t(a_{\text{mr}})$ and comparing their values.

Solution to Exercise 7.2

(a) The question tells us to assume that $\Omega_r \gg \Omega_b$. This means that $\Omega_b/\Omega_r \ll 1$, and so the expression for the speed of sound in Equation 7.5 can be simplified to

$$c_s = c \left(3 + \frac{9}{4} \frac{\Omega_b}{\Omega_r} \right)^{-1/2} \approx \frac{c}{\sqrt{3}}$$

Substituting this approximation in Equation 7.4 and using the expression for the horizon distance in Equation 7.1 yields the required expression for the maximum distance an acoustic oscillation could have travelled by the time of last scattering:

$$\begin{aligned} d_s(z_{\text{ls}}) &= a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{c_s}{a(t)} dt \\ &\approx a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{c}{\sqrt{3} a(t)} dt \\ &\approx \frac{c}{\sqrt{3}} a(z_{\text{ls}}) \int_0^{t_{\text{ls}}} \frac{dt}{a(t)} \\ &\approx \frac{d_{\text{hor}}(z_{\text{ls}})}{\sqrt{3}} \end{aligned}$$

(b) Now we can calculate the acoustic scale using Equation 7.6:

$$\begin{aligned} \theta_s &\approx \frac{d_s(z_{\text{ls}})}{d_A(z_{\text{ls}})} \approx \frac{d_{\text{hor}}(z_{\text{ls}})}{\sqrt{3} d_A(z_{\text{ls}})} \\ &\approx \frac{0.29 \text{ Mpc}}{\sqrt{3} \times 12.73 \text{ Mpc}} = 0.013 \text{ rad} \approx 0.75^\circ \end{aligned}$$

As you learned in Section 6.3.1, fluctuations with angular scale $\Delta\theta$ radians correspond to multipole number $\sim \pi/\ell$. Therefore, because π radians is equivalent to 180 degrees, the angular scale represented by θ_s corresponds to multipole number

$$\ell_s \approx \frac{180^\circ}{0.75^\circ} \approx 240$$

which is remarkably close to the observed location of the first acoustic peak in Figure 7.1.

Solution to Exercise 8.1

Your completed table should look similar to Table S2. As noted in the question, you may have found somewhat different numbers for the relative strengths of the forces. Those quoted here are the relative strength of the four forces acting between two protons that are just in contact.

Table S2 Summary of the basic properties of the four forces.

Interaction	Particles affected	Carrier particle	Relative strength
Strong	quarks	gluon	1
Weak	quarks, baryons, leptons	W and Z bosons	10^{-2}
Electromagnetic	all charged particles, leptons	photon	10^{-7}
Gravity	all massive particles	none (Standard Model)	10^{-39}

Solution to Exercise 8.2

The interaction energy in J is

$$E_{\text{int}} \approx 1.381 \times 10^{-23} \text{ J K}^{-1} \times 10^{14} \text{ K}$$

$$\approx 1.381 \times 10^{-9} \text{ J}$$

Converting to GeV:

$$\frac{1.381 \times 10^{-9} \text{ J}}{1.602 \times 10^{-19} \text{ J eV}^{-1}} = 8.620 \times 10^9 \text{ eV}$$

$$\approx 8.6 \text{ GeV}$$

Solution to Exercise 8.3

As you read in Section 8.1.1, in order to create particles the interaction energy must be higher than the total rest mass energy of the corresponding particles, as given by $E = mc^2$. Therefore:

- a proton–antiproton pair requires energy of $E = 2m_p c^2 = 3.0 \times 10^{-10} \text{ J} = 1.9 \text{ GeV}$
- an electron–positron pair requires $E = 2m_e c^2 = 1.6 \times 10^{-13} \text{ J} = 0.001 \text{ GeV}$.

So, in terms of the interaction energy required, it is much easier to create electron–positron pairs than proton–antiproton pairs.

Solution to Exercise 8.4

Using Equation 8.15 and substituting in the relevant particle values, converting to appropriate units and recalling that the interaction energy E is approximately equal to $k_B T$ (Equation 8.2) gives:

$$\begin{aligned} \frac{n_n}{n_p} &= \exp \left[-\frac{(m_n - m_p)c^2}{k_B T} \right] \\ &= \exp \left(-\frac{2.067 \times 10^{-13} \text{ kg m}^2 \text{ s}^{-2}}{1.282 \times 10^{-13} \text{ kg m}^2 \text{ s}^{-2}} \right) \\ &= 0.20 \end{aligned}$$

In other words, there are five protons for every neutron, as required.

Solution to Exercise 8.5

We can rearrange Equation 8.16 to find an expression for t :

$$\frac{N(t)}{N(t_{\text{init}})} = \exp(-t/\tau_n)$$

and so

$$t = -\tau_n \ln[N(t)/N(t_{\text{init}})]$$

Because we are told that n_p can be approximated as remaining constant (because $n_p \gg n_n$ when the decay starts), the ratio of the number of neutrons is the same as the ratio between n_n/n_p at the two times:

$$\frac{N(t)}{N(t_{\text{init}})} = \frac{n_n/n_p(t)}{n_n/n_p(t_{\text{init}})}$$

Therefore t can be obtained from

$$t = -\tau_n \ln \left(\frac{n_n/n_p(t)}{n_n/n_p(t_{\text{init}})} \right)$$

Substituting in the provided values gives $t = 1230 \text{ s}$. Because t is the time elapsed since $t_{\text{init}} = 1 \text{ s}$, a more precise estimate of the time since the big bang at which this neutron-to-proton ratio would apply is 1231 s.

Solution to Exercise 8.6

Please refer to the Exercise 8.6 Jupyter Notebook solution available from the online module resources to see the solution to this example.

The most likely pitfall when attempting this task is in setting the integration ranges – this requires converting from photon energy to frequency, and using the provided hint to obtain a maximum frequency for the calculation of the total number of photons.

Solution to Exercise 9.1

We can make use of the final equation obtained in Example 9.1 (Equation 9.5) to answer this question. It is not trivial to solve this expression algebraically, but we can simply substitute values into the left- and right-hand sides and demonstrate equality.

Taking the definition of recombination corresponding to $X = 0.5$, the left-hand side of Equation 9.5 becomes:

$$\frac{1-X}{X^2} = 2.0$$

Substituting the provided temperature and value of η into the right-hand side gives:

$$10^{-9} \left[\frac{a(3795)^3}{3k_B} \right] \times \left[\frac{m_e k_B (3795)}{2\pi\hbar^2} \right]^{-3/2} \times \exp \left[\frac{Q}{k_B (3795)} \right] = 2.00$$

Therefore the left- and right-hand sides are in agreement (to within a few per cent) for $T = 3795$ K.

Solution to Exercise 9.2

The two measured D/H values of 2.52×10^{-5} (quasars) and 1.38×10^{-5} (Milky Way), correspond to the abundances predicted by big bang nucleosynthesis for $\eta \sim 6 \times 10^{-10}$ and $\sim 9 \times 10^{-10}$, respectively, according to Figure 9.3. (Note that these are necessarily fairly rough estimates given the precision of the graph.)

We can convert these to values of $\Omega_{b,0}$ using Equation 8.11 from Chapter 8, which rearranges

to give:

$$\Omega_{b,0} = \eta \frac{\langle m \rangle n_{\gamma,0}}{\rho_{c,0}}$$

If $\langle m \rangle = 1.6726 \times 10^{-27}$ kg, $n_{\gamma,0} = 4.0 \times 10^8$ m⁻³ and $\rho_{c,0} = 8.599 \times 10^{-27}$ kg m⁻³, this gives $\Omega_{b,0} \approx 0.05$ (quasars) and $\Omega_{b,0} \approx 0.07$ (Milky Way). Hence the quasar-measured estimate of η is in much better agreement with the baryon density parameter measured via the CMB (given as 0.0490 in the table of constants) than the Milky Way estimate.

Solution to Exercise 9.3

If the baryon-to-photon ratio is low then the start of deuterium production is delayed, because the high-energy tail of the photon distribution is more successful at dissociating the deuterons as they form. The consequence of this deuterium bottleneck is that the ratio of neutrons to protons continues to decrease, so there are fewer neutrons available for nucleosynthesis when η is low compared to when it is high.

The vast majority of deuterium produced in the BBN period goes on to form helium via the reactions shown in Section 8.3.2, so more helium forms for high η (when lots of neutrons are available) than for low η . The decreasing abundance of deuterium occurs because the reactions that create helium are more efficient when there are more neutrons present, so a greater proportion of the deuterium is converted to helium.

Solution to Exercise 9.4

(a) The total mass can be calculated by substituting the provided velocity dispersion and radius into Equation 9.18, which gives $M = 4.84 \times 10^{45}$ kg = 2.43×10^{15} M_⊕.

(b) The mass-to-light ratio is simply the total mass, in units of solar mass, divided by the optical luminosity, in units of solar luminosity. In other words $M/L = (2.43 \times 10^{15} \text{ M}_{\odot}) / (5.0 \times 10^{12}) \approx 490$.

(c) By definition, the Sun has a mass-to-light ratio of 1 when considered in units of M_⊕ and L_⊕. Therefore, a system with $M/L \gg 1$ has proportionally much more mass for the same

Solutions to exercises

amount of light it is producing. If the Coma cluster's mass was primarily made up of stars, then on average the stars would need to be >500 times less luminous than the Sun. This is not realistic.

Solution to Exercise 9.5

Rearranging Equation 9.20 for mass gives:

$$M = \frac{5k_BRT}{G\bar{m}}$$

Substituting temperatures of 10^7 K and 10^8 K gives a mass range of 3.2×10^{44} – 3.2×10^{45} kg, which corresponds to 1.6×10^{14} – 1.6×10^{15} M_⊕.

Solution to Exercise 10.1

The gravitational acceleration of the two galaxies would be:

$$\begin{aligned} g &= \frac{GM}{r^2} = \frac{G \times 10^{12} \text{ M}_\odot}{(3 \text{ Mpc})^2} \\ &= \frac{6.673 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2} \times 1.99 \times 10^{42} \text{ kg}}{(3 \times 3.086 \times 10^{22} \text{ m})^2} \\ &= 1.55 \times 10^{-14} \text{ m s}^{-2} \end{aligned}$$

It would therefore take

$$\frac{200 \text{ km s}^{-1}}{1.55 \times 10^{-14} \text{ m s}^{-2}} = 1.29 \times 10^{19} \text{ s}$$

or about 30 times the current age of the Universe to bring them to rest. Given that H_0 appears to be increasing (i.e. the expansion is accelerating), it seems these two galaxies will never come together.

Solution to Exercise 10.2

The mean density for each object can be obtained by assuming a sphere of radius, R , so that

$$\rho = \frac{M}{\frac{4}{3}\pi R^3}$$

The free-fall timescale can then be obtained via Equation 10.6:

$$t_{\text{ff}} \approx 18.45 \text{ hours} \left(\frac{\rho}{1 \text{ kg m}^{-3}} \right)^{-1/2}$$

The resulting values for each object are given in Table S3.

Table S3 Densities and free-fall timescales for the Milky Way and the Local Group.

System	$\rho/\text{kg m}^{-3}$	t_{ff}
Milky Way	1.4×10^{-21}	57 My
Local Group	6.0×10^{-26}	8.6 Gy

These results show that objects of the density of the present-day Milky Way can collapse on a relatively short cosmological timescale if unopposed: 57 million years is less than the time since the demise of the dinosaurs! By contrast, a halo of the mean density of the Local Group would require a substantial fraction (about half) of the age of the Universe to collapse.

This calculation started from the densities of the present-day (collapsed) objects. The collapse timescales to form these objects starting from the less dense conditions of a primordial cloud would, in principle, be longer.

Solution to Exercise 10.3

Equation 10.16 states $\delta(t) \propto 1/(1+z)$.

The question states that at the time of decoupling, $z \approx 1090$, so $1/(1+z) \approx 9.17 \times 10^{-4}$.

At $z \approx 13$, we have $1/(1+z) \approx 7.14 \times 10^{-2}$.

If $\delta \approx 10^{-5}$ at decoupling, as implied by fluctuations in the CMB temperature, then δ will have only grown to $(10^5) \times (7.14 \times 10^{-2}/9.17 \times 10^{-4}) \sim 8 \times 10^{-4}$ by $z \approx 13$, and $\delta \approx 10^{-2}$ today. Galaxies have densities many times higher than the mean density of the Universe, i.e. $\delta > 1$. Although Equation 10.16 is invalid as δ approaches 1, this exercise shows it should not yet have done so. In other words, Equation 10.16 implies that galaxies should not exist today, never mind at $z \approx 13$.

The solution to this apparent inconsistency is to recall that the density fluctuations in the CMB describe the behaviour of the photon–baryon fluid, not the dark matter. In order for galaxy-sized overdensities to form, the dark-matter overdensities at the time the CMB was produced must have been larger by a substantial factor than those of the baryons and photons.

Solution to Exercise 11.1

- (a) The gas is able to cool over the density range where the track has a negative slope (i.e. temperature is decreasing as density increases); this occurs between the edge of the diagram at $n = 1 \text{ cm}^{-3}$ and $n \approx 10^5 \text{ cm}^{-3}$. Over the rest of the density range (i.e. where $n > 10^5 \text{ cm}^{-3}$) the gas heats as it becomes denser, and therefore cannot fragment into smaller clumps.
- (b) The solar-composition line reaches a minimum just to the right of the line corresponding to a Jeans mass of $1 M_{\odot}$. This is the optimal place for star formation, as it has the highest density and lowest temperature. It suggests that typical stars would have masses a little less than that of the Sun, which is indeed what is observed in the Milky Way.
- (c) Considering the $Z = 0$ cooling track, which corresponds to the expected behaviour of primordial gas, the minimum occurs at a much higher Jeans mass than the solar-composition line ($> 100 M_{\odot}$). This is because the cooling mechanisms involving emission lines are very much less effective at low metallicity. The first generation of stars were therefore probably much more massive than stars today, with an average mass of several hundred solar masses.

Solution to Exercise 11.2

- (a) The cooling timescale can be estimated from Figure 11.3. For a gas cloud with $Z = 0$ and $T \approx 3000 \text{ K}$ and no molecular hydrogen, Figure 11.3 indicates that $\Lambda_{\text{cool}} \approx 3 \times 10^{-29} \text{ J m}^{-3} \text{ s}^{-1}$. Using this value in Equation 11.1 with $\gamma = 3$ (assuming monatomic hydrogen) gives:

$$\begin{aligned} t_{\text{cool}} &= \frac{3nk_B T}{2\Lambda_{\text{cool}}} \\ &\approx \frac{3 \times 3 \times 10^8 \text{ m}^{-3} \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 3000 \text{ K}}{2 \times 3 \times 10^{-29} \text{ J m}^{-3} \text{ s}^{-1}} \\ &\approx 6.2 \times 10^{17} \text{ s} \\ &\approx 20 \text{ Gy} \end{aligned}$$

We have assumed that the gas started cooling from 3000 K at around the time of recombination, t_{rec} ,

and so the age of the Universe at which stars first formed is $t_{\text{cool}} + t_{\text{rec}}$. In this case the timescale is dominated by the cooling time, and so the age of the Universe when stars first formed would still be 20 Gy.

- (b) For a gas cloud with 1% molecular hydrogen, when $T \approx 3000 \text{ K}$ then $\Lambda_{\text{cool}} \approx 3 \times 10^{-25} \text{ J m}^{-3} \text{ s}^{-1}$. Using this value in Equation 11.1, this time with $\gamma = 6$ for a molecular gas, gives:

$$\begin{aligned} t_{\text{cool}} &= \frac{6nk_B T}{2\Lambda_{\text{cool}}} \\ &\approx \frac{6 \times 3 \times 10^8 \text{ m}^{-3} \times 1.381 \times 10^{-23} \text{ J K}^{-1} \times 3000 \text{ K}}{2 \times 3 \times 10^{-25} \text{ J m}^{-3} \text{ s}^{-1}} \\ &\approx 1.2 \times 10^{14} \text{ s} \\ &\approx 4 \text{ My} \end{aligned}$$

In this case the age of the Universe at which stars first formed is $t_{\text{cool}} + t_{\text{rec}} \approx 4.4 \text{ My}$.

- (c) The age of the Universe when stars first formed calculated in part (a) for the zero metallicity case is considerably greater than the actual age of the Universe according to current models and observations ($\approx 13.7 \text{ Gy}$). This provides strong evidence that molecular hydrogen cooling must be important to enable the first stars to form: without it there would not yet be stars, and we wouldn't be here to think about them. The age calculated in part (b) for the 1% H_2 scenario is quite short: in reality it is thought that the first stars took around 200 million years to form, and so although we have shown that molecular hydrogen must be important, more sophisticated modelling is needed to obtain a more accurate timescale for the first stars.

- (d) The main simplification that has been made in our calculations is to ignore the fact that the parameters of Equation 11.1 evolve:

- the baryons initially continue to expand and dilute with the Universe, so n goes down
- as a cloud of baryons collapses into a star, n goes up
- the gas is still being heated by the CMB, so T stays high
- the cooling rate, Λ_{cool} , will change as the gas cools (and as it becomes denser).

Solutions to exercises

Accounting for the evolution of these parameters, and more accurate modelling of the H₂ fraction, would enable a more accurate timescale to be obtained.

Solution to Exercise 11.3

(a) We first note that

$$\begin{aligned} 10 \text{ cm}^{-3} &= 10 \times (10^{-2} \text{ m})^{-3} \\ &= 10^7 \text{ m}^{-3} \end{aligned}$$

Then for scenario (i) the maximum radius is:

$$\begin{aligned} R_{\max} &\approx \left[\frac{3Q}{4\pi n_i n_e \alpha(T)} \right]^{1/3} \\ &\approx \left[\frac{3 \times 10^{50} \text{ s}^{-1}}{4\pi \times (10^7 \text{ m}^{-3})^2 \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1}} \right]^{1/3} \\ &\approx 8.42 \times 10^{17} \text{ m} \\ &\approx 27 \text{ pc} \end{aligned}$$

For scenario (ii) first note that:

$$\begin{aligned} 10^{-5} \text{ cm}^{-3} &= 10^{-6} \times (10^{-2} \text{ cm})^{-3} \\ &= 10 \text{ m}^{-3} \end{aligned}$$

Then the maximum radius is:

$$\begin{aligned} R_{\max} &\approx \left[\frac{3 \times 10^{56} \text{ s}^{-1}}{4\pi \times (10 \text{ m}^{-3})^2 \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1}} \right]^{1/3} \\ &\approx 2.71 \times 10^{23} \text{ m} \\ &\approx 8.8 \text{ Mpc} \end{aligned}$$

(b) For scenario (i) the recombination timescale is:

$$\begin{aligned} t_r &= (10^7 \text{ m}^{-3} \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1})^{-1} \\ &= 2.5 \times 10^{11} \text{ s} \\ &= 8000 \text{ y} \end{aligned}$$

while for scenario (ii):

$$\begin{aligned} t_r &= (10 \text{ m}^{-3} \times 4 \times 10^{-19} \text{ m}^3 \text{ s}^{-1})^{-1} \\ &= 2.5 \times 10^{17} \text{ s} \\ &\approx 7.9 \times 10^9 \text{ y} \end{aligned}$$

(c) For scenario (i) the size of the Strömgren sphere after 10⁶ years is:

$$\begin{aligned} R(t) &\approx R_{\max} [1 - \exp(-t/t_r)]^{1/3} \\ &\approx 27 \text{ pc} [1 - \exp(-10^6/8000)]^{1/3} \\ &\approx 27 \text{ pc} \end{aligned}$$

and for scenario (ii):

$$\begin{aligned} R(t) &\approx 8.8 \text{ Mpc} [1 - \exp(-10^6/(7.9 \times 10^9))]^{1/3} \\ &\approx 0.44 \text{ Mpc} \end{aligned}$$

[For interest, this is around half the distance light can travel in that time. Consequently, we expect the Strömgren sphere to expand out of this (small) galaxy at an appreciable fraction of the speed of light.]

Solution to Exercise 11.4

(a) Using the numbers given, the escape speed is

$$\begin{aligned} v_{\text{esc}} &\approx \sqrt{\frac{2G \times 10^{12} \text{ M}_\odot}{8 \text{ kpc}}} \\ &\approx 1000 \text{ km s}^{-1} \end{aligned}$$

(b) The escape speed is (i) slightly greater than that of the supernova, but (ii) much greater than that of the low-mass star.

(c) Supernovae ejecta will therefore not escape from the galaxy, but will be flung out far into the galaxy's halo before falling back down. Ejecta from dying low-mass stars will contribute directly to its local environment, cycling matter directly back into the galaxy's interstellar medium.

(d) For a galaxy of $M = 10^8 \text{ M}_\odot$ and $R = 800 \text{ pc}$, the escape speed will only be $v_{\text{esc}} \approx 33 \text{ km s}^{-1}$. Consequently, while the galaxy will still retain the ejecta of its low-mass stars, it will lose the ejecta of its supernovae. This will lead to a different overall chemical composition for the smaller galaxy as its retained ejecta will be enriched only in the elements created by dying low-mass stars, which can be seen in Figure 11.8 to differ from those produced in supernovae (exploding massive stars).

Solution to Exercise 11.5

- (a) As indicated in Figure 11.8, hydrogen is formed in the big bang, oxygen is formed from exploding massive stars, and (although some iron is produced by exploding massive stars) most iron is produced by exploding white dwarfs.
- (b) To begin with, the Universe is composed of hydrogen and helium. Recalling that time evolution proceeds from left to right in Figure 11.7, massive stars will be the first to die. However, for white dwarfs to reach the point where they inject metals into the ISM, less-massive stars have to go through their entire (longer) evolution to produce white dwarfs. The white dwarfs then have to evolve further until they become Type Ia supernovae.
- (c) As massive stars die, the oxygen and iron abundances will increase in tandem, creating a constant [O/Fe] ratio and increasing [Fe/H] ratio. As white dwarfs explode and contribute more iron, the [Fe/H] ratio will continue to increase, but the extra iron will make the [O/Fe] ratio decrease.
- (d) The different points where the [O/Fe] ratio decreases mark the different levels of enrichment each galaxy achieved by the time exploding white dwarfs started to contribute. (While the first white dwarfs form from intermediate-mass stars in 35 My, the process of chemical enrichment takes much longer, as they then need to accrete enough matter to become Type Ia supernovae.)

Solution to Exercise 11.6

- (a) In order to predict the number density of galaxies in a particular stellar mass range we first need to identify the corresponding halo mass range for each case. The question tells us to assume $f_B = 0.15$ and $\epsilon_* = 1$. We can therefore assume that all of the baryons in the corresponding halos are now in the form of stars, and so the total mass of baryons in each halo is the same as the stellar mass. This means that, according to Equation 11.14, the corresponding halo mass range must be $1/0.15 \approx 6.7$ times the stellar mass range.

The relevant mass ranges to consider are therefore:

- (i) $6.7 \times 10^8 < M_{\text{halo}} < 6.7 \times 10^9 M_\odot$
- (ii) $6.7 \times 10^{10} < M_{\text{halo}} < 6.7 \times 10^{11} M_\odot$
- (iii) $2.1 \times 10^{12} < M_{\text{halo}} < 2.1 \times 10^{13} M_\odot$.

We can now determine the number density of halos using the same method as for Example 11.1. Substituting each of the mass ranges into Equation 11.13 gives galaxy densities of

- (i) $n = 1.9 \text{ galaxies Mpc}^{-3}$
 - (ii) $n = 0.03 \text{ galaxies Mpc}^{-3}$
 - (iii) $n = 0.001 \text{ galaxies Mpc}^{-3}$.
- (b) Estimating the observed galaxy number densities for the middle of each *stellar mass range* from Figure 11.13 gives a little above ~ 0.03 , ~ 0.01 and $\sim 10^{-5}$, respectively. Therefore, the assumptions of constant baryon fraction and star-formation efficiency cannot be correct. The prediction is closest for the intermediate mass range, but predicts far too many galaxies for both the low and high stellar mass ranges.

References and acknowledgements

References

- Bernardi, M. *et al.* (2013) ‘The massive end of the luminosity and stellar mass functions: dependence on the fit to the light profile’, *Monthly Notices of the Royal Astronomical Society*, 436(1), pp. 697–704. Available at <https://doi.org/10.1093/mnras/stt1607>.
- Cooke, R. J. *et al.* (2014) ‘Precision measurements of the primordial abundance of deuterium’, *The Astrophysical Journal*, 781(1), 1. Available at <https://doi.org/10.1088/0004-637X/781/1/31>.
- Cyburt, R. H. *et al.* (2016) ‘Big bang nucleosynthesis: present status’, *Reviews of Modern Physics*, 88(1), 015004. Available at <https://doi.org/10.1103/RevModPhys.88.015004>.
- Lellouch, E. *et al.* (2001) ‘The deuterium abundance in Jupiter and Saturn from ISO-SWS observations’, *Astronomy & Astrophysics*, 370(2), pp. 610–622. Available at <https://doi.org/10.1051/0004-6361:20010259>.
- Linsky, J. L. *et al.* (2006) ‘What is the total deuterium abundance in the local Galactic disk?’, *The Astrophysical Journal*, 647(2), pp. 1106–1124. Available at <https://doi.org/10.1086/505556>.
- Ryden, B. (2017) *Introduction to cosmology*. 2nd edn. New York: Cambridge University Press.
- Sbordone, L. *et al.* (2010) ‘The metal-poor end of the Spite plateau’, *Astronomy & Astrophysics*, 522, A26. Available at <https://doi.org/10.1051/0004-6361/200913282>.
- Wright, A. H. *et al.* (2017) ‘Galaxy And Mass Assembly (GAMA): the galaxy stellar mass function to $z = 0.1$ from the r -band selected equatorial regions’, *Monthly Notices of the Royal Astronomical Society*, 470(1), pp. 283–302. Available at <https://doi.org/10.1093/mnras/stx1149>.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: Max-Planck Institute for Physics.

Chapter images: Figure 6.2a: BICEP/Keck Collaboration; Figure 6.2b: Amble, https://commons.wikimedia.org/wiki/File:South_pole_telescope_nov2009.jpg, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported (CC BY SA 3.0) license, <https://creativecommons.org/licenses/by-sa/3.0/deed.en>; Figure 6.2c: Nathan Precup; Figure 6.3: NASA/JPL-Caltech/ESA; Figures 6.5, 6.7 and 6.9: this research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology; Figure 6.8: ESO, <https://arxiv.org/abs/1502.01588>, licensed under a Creative Commons Attribution 4.0 International (CC BY SA 3.0) license, <https://creativecommons.org/>

References and acknowledgements

licenses/by/4.0/; Figure 7.12: adapted from Suzuki, N. *et al.* (2012) ‘The Hubble Space Telescope cluster supernova survey. V. Improving the dark-energy constraints above $z > 1$ and building an early-type-hosted supernova sample’, *The Astrophysical Journal*, 746(1), p. 85, The American Astronomical Society; Figure 9.2: Sloan Digital Sky Survey; Figure 9.6: Izotov, Y. I. *et al.* (2014) ‘A new determination of the primordial He abundance using the HeI $\lambda 10830 \text{ \AA}$ emission line: cosmological implications’, *Monthly Notices of the Royal Astronomical Society*, 445(1), pp. 778–793, Royal Astronomical Society; Figure 9.7: adapted from Figure 15 of Sbordone, L., *et al.* (2010) ‘The metal-poor end of the Spite plateau’, *Astronomy & Astrophysics* 522, A26, EDP Sciences, https://www.aanda.org/articles/aa/full_html/2010/14/aa13282-09/F15.html; Figure 9.9a: NASA/STScI; Figure 9.9b: NASA/CXC/MIT/Peng, E.-H. *et al.*; Figure 9.10: NASA/STScI, Magellan/U. Arizona/Clowe, D. *et al.*; Figure 9.11a: Justin Yaros and Andy Schleif/Flynn Haase/NOAO/AURA/NSF; Figure 9.11b: adapted from Begeman, K. G. *et al.* (1991) ‘Extended rotation curves of spiral galaxies: dark haloes and modified dynamics’, *Monthly Notices of the Royal Astronomical Society*, 249(3), pp. 523–537; Figure 9.12: Sofue, Y. *et al.* (2009) ‘Unified rotation curve of the Galaxy – decomposition into de Vaucouleurs bulge, disk, dark halo, and the 9-kpc rotation dip’, *Astronomical Society of Japan*, 61(2), Astronomical Society of Japan; Figure 10.2a: Bender, R. IMPRS astrophysics introductory course, Lecture 7; Figure 10.4: Illustris Collaboration/Illustris Simulation; Figure 10.5: Springel, V. Max Planck Institute for Astrophysics; Figure 10.6: Vikhlinin, A. *et al.* (2009) ‘Chandra cluster cosmology project III: cosmological parameter constraints’, *The Astrophysical Journal* 692(2), IOP Publishing; Figure 11.2: The THESAN Collaboration/HTML5 UP, https://www.thesan-project.com/images/media/thesan_lightcone.png, this work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.4: Klessen, R. S. *et al* (2012) ‘On the formation of very metal poor stars: the case of SDSS J1029151+172927’, *Monthly Notices of the Royal Astronomical Society*, 421(4), pp. 3217–3221, Oxford University Press; Figure 11.5a, b and c: Marcelo Alvarez (CITA), Ralf Kaehler (Stanford), Tom Abel (Stanford); Figure 11.8: Cmglee, https://commons.wikimedia.org/wiki/File:Nucleosynthesis_periodic_table.svg, this file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0) license, <https://creativecommons.org/licenses/by-sa/3.0/deed.en>; Figure 11.9a: NASA/CXC/SAO; Figure 11.9b: NASA, ESA and the Hubble Heritage Team (STScI/AURA), Gallagher, J. (University of Wisconsin), Mountain, M. (STScI) and Puxley, P. (NSF), <https://esahubble.org/images/heic0604a/>, released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.9c: NASA Goddard Space Flight Center; Figure 11.9d: NASA/ESA/STScI/AURA/The Hubble Heritage; Figure 11.10a: NRAO/AUI/NSF/

Univ. Hertfordshire/Hardcastle, M. <https://public.nrao.edu/gallery/centaurus-a-in-radio/>, licensed for use under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, <https://creativecommons.org/licenses/by/3.0/>; Figure 11.10b: ESO, <https://www.eso.org/public/images/eso1221a/>, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Figure 11.10c: ESO/WFI (optical), MPIfR/ESO/APEX/Weiss, A. *et al.* (submillimetre), NASA/CXC/CfA/Kraft, R. *et al.* (X-ray), <https://www.eso.org/public/images/eso0903a/>, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>; Madau, P. and Dickinson, M. (2014) ‘Cosmic star-formation history’ *Annual Review of Astronomy and Astrophysics*, 52, pp. 415–486. Available at <https://doi.org/10.1146/annurev-astro-081811-125615>; Figure 11.13: Bullock, J. S. and Boylan-Kolchin, M. (2017) ‘Small-scale challenges to the Λ CDM Paradigm’, *Annual Review of Astronomy and Astrophysics*, 55, pp. 343—387. Available at <https://doi.org/10.1146/annurev-astro-091916-055313>; Figure 11.14: Sánchez, A. G. *et al.* (2012) ‘The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological implications of the large-scale two-point correlation function’, *Monthly Notices of the Royal Astronomical Society*, 425(1), pp. 415–437. Available at <https://doi.org/10.1111/j.1365-2966.2012.21502.x>; Figure 11.15: Eisenstein, D. J. *et al.* (2006) ‘On the robustness of the acoustic scale in the low-redshift clustering of matter’, *The Astrophysics Journal*, 664, pp. 660–674, IOP Publishing, American Astronomical Society, Institute of Physics, University of Chicago Press.

Software: ‘Python’ and the Python logos are trademarks or registered trademarks of the Python Software Foundation, used by The Open University with permission from the Foundation.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Book production contributors

Academic authors

Judith Croston (Chair), Hugh Dickinson, Iain McDonald and Sheona Urquhart.

The authors would like to thank Stephen Serjeant, Bonny Barkus, Kate Gibson and Mark Jones for useful feedback and discussions.

External assessor

Stephen Wilkins, University of Sussex.

Curriculum team

Jessica Bartlett and Shelah Surveye.

Production team

Senior project manager

Jeni Aldridge.

Editors

Jonathan Martyn, Peter Twomey, Yon-Hee Kim and Lil Davies.
Mark Radford (Pepperhouse Editorial) and Jonathan Darch.

Graphics

Sha'ni Hirschy.

OU Library

James Salter.

Index

Note: **bold** page numbers indicate where terms are defined.

Abell 1689 99
 absorption line 89
 abundance 88
 abundance ratio 90
 acoustic oscillations **6**, 33
 acoustic peak **35**, 41, 44, 46, 50
 acoustic scale **42**, 44
 active galactic nucleus (AGN) **141**, 146
 AGN *see* active galactic nucleus
 AGN feedback **146**
 Aitoff projection **19**
 Andromeda 110
 angular power spectrum **23**, 25, 30
 annihilation **62**, 70, 72
 antibaryon 62
 antilepton 62
 antiquark 62
 archaeology
 galactic 147
 associated Legendre polynomial **19**
 astration **91**, 94

 BAOs *see* baryon acoustic oscillations
 baryogenesis 70
 baryon acoustic oscillations (BAOs) 54, **155**
 baryon number **62**
 baryon-to-photon ratio **71**, 84
 baryonic matter 31
 BBN *see* big bang nucleosynthesis
 beta decay **63**
 BICEP Array 6
 BICEP3 6
 big bang nucleosynthesis (BBN) **65**, 74, 91
 binary star 145
 black hole 62, 141, 145
 supermassive 146
 Boltzmann factor **73**
 bound halo **120**
 bremsstrahlung radiation **13**
 Bullet cluster 103

 Centaurus A 146
 chemical enrichment **90**, 141
 clustering of galaxies 154

CMB *see* cosmic microwave background
 CMB anisotropies **6**
 CMB rest frame 10
 CMB solar dipole **9**
COBE *see* *Cosmic Background Explorer*
 cold dark matter **107**
 Coma cluster 99, 102
 Compton scattering **65**
 cooling rate **133**
 cooling timescale **135**
 correlation function
 two-point 154
Cosmic Background Explorer (COBE) 7, 8
 cosmic cycle of matter 141
 cosmic microwave background (CMB) 3
 angular power spectrum 23, 25, 30
 anisotropies 6
 energy density 40
 rest frame 10
 solar dipole 9
 temperature 6, 71
 temperature fluctuation 15
 temperature fluctuation amplitude 18
 cosmic neutrino background 70
 cosmic variance **24**
 cosmic web 122
 cosmological constant 31, 50
 curvature of Universe 44

 dark ages 131, 132
 dark matter 31, 58, 99
 cold 107
 evidence 103, 106
 hot 107
 dark-matter halo 120, 130
 decouple **61**, 72, 87
 degenerate parameter **53**
 density parameter 31, 33, 42, 44, 46, 49, 50
 deuterium 65, 77
 deuterium abundance 91, 94
 deuterium bottleneck **76**, 79
 deuterium formation 74
 deuteron 65, 74
 deuteron formation 65

- dissociation 63
Doppler peak *see* acoustic peak
Doppler shift 10, 40, 101
electric dipole moment **13**
electromagnetic interaction 59
electron capture **64**
electroweak transition 69
electroweak unification **68**
elemental abundance **88**
emission line 89
enrichment 90, 141
epoch of last scattering 4, 31
epoch of reionisation **139**
estimators **23**
expansion of Universe 38, 58, 115, 156
faint-end slope **152**
filament **121**
forbidden lines **132**
Fourier expansion **18**
free-fall timescale **114**
free-free emission 13
freeze out **61**
galactic archaeology **147**
Galactic foreground emission 13
galactic outflow **143**
galaxy
 formation 140
 minimum mass 145
galaxy cluster **99**
 Abell 1689 99
 Bullet 103
 Coma 99, 102
 formation 154
 weight 99
galaxy merger 103, **141**
giant molecular clouds 13
grand unification 68
grand unified theory (GUT) **68**
gravitational lensing 103
gravitational potential well 31, 35
gravitational waves 31
grid simulation **123**
GUT *see* grand unified theory
H₂ formation heating **134**
hadron era 70
hadron formation 63
halo mass function **125**, 151, 154
helium abundance 78, 94, 95
helium formation 77
helium-3 77
helium-4 77, 80
HII regions **96**
hot dark matter 107
Hubble constant 46
Hubble diagram 110
Hubble flow 110
Hubble–Lemaître relationship 110
Hubble parameter 61, 87, 156
IGM *see* intergalactic medium
Illustris simulation 124
inertial frame 9
inflation **68**
interacting binary star **145**
interaction cross-section **59**
interaction energy **60**
intergalactic medium (IGM) **138**
interstellar dust 13, 134
interstellar medium (ISM) **13**, 91, 136, 141
inverse Compton scattering **65**
ionisation 67
ionisation fraction 84
ISM *see* interstellar medium
jansky (Jy) 7
Jeans criterion **111**, 112
Jeans length **111**
Jeans mass **111**, 135
jet 141, **146**
Jupiter 94
Keplerian rotation **105**
Large Magellanic Cloud (LMC) 148
last scattering *see* epoch of last scattering,
 surface of last scattering
Legendre polynomial 19
 evaluation 21
lepton number **62**
lithium 80
lithium abundance 97
LMC *see* Large Magellanic Cloud
Local Group **10**, 110, 114
Lorentz transformations 10

- luminosity function **150**, 154
 Lyman limit **136**
 major merger 141
 mass deficit **66**
 mass fraction **78**, 90
 mass segregation **145**
 merger
 galaxy 103, 141
 stellar 65
 Messier 31 110
 Messier 82 143
 metallicity **90**, 96
 metals 90, 133
 Milky Way 10, 17, 92–94, 110, 114, 148
 total mass 106
 Millennium Simulation 125
 minor merger 141
 moving-mesh simulation **123**
 multipole number 19
N-body simulation **123**
 nested sampling 123
 neutrino decoupling 70
 neutron star 145
 neutron-to-proton ratio 72, 78
 Newton's shell theorems 113
 NGC 1560 105
 NGC 5128 146
 nucleosynthesis *see* big bang nucleosynthesis
 number counts **149**
 numerical simulation 122, *see also* simulation
 overdensity 10, 31, 35, 39, 112, 114, 118, 130,
 156
 pair production **61**, 62
 parameter degeneracies **53**
 photodisintegration **66**, 74
 photoionisation 67
 photon–baryon fluid **33**, 33–36, 40, 155
Planck 7–9, 16, 91, 92
 Planck time **68**
 Population I **136**
 Population II **136**, 142
 Population III **136**, 142
 positron capture **64**
 primordial nucleosynthesis **74**, *see* big bang
 nucleosynthesis
 protogalaxy **140**
 quasar 91
 quenching **146**
 radiation driving **39**, 49
 radio lobe 141, **146**
 reaction rate **59**
 recombination 67, 83–88, 132, 137–138
 recombination coefficient **137**
 red and dead 145
 red giant 142
 reionisation **136**, 139
 relative abundance **90**
 RMS *see* root mean square
 root mean square (RMS) **14**
 rotation curve 104
 Sachs–Wolfe effect **39**
 Sachs–Wolfe plateau **32**
 Sagittarius 148
 Saha equation 67, 84
 Saturn 94
 scale factor 68, 115, 161
 Sculptor 148
 SDSS *see* Sloan Digital Sky Survey
 shell theorems 113
 Silk damping **40**
 simulation 107, 122–126, 139
 Illustris 124
 Millennium Simulation 125
 types 122
 Sloan Digital Sky Survey (SDSS) 89
 SMBH *see* supermassive black hole
 smoothed-particle hydrodynamic (SPH)
 simulation **123**
 sound horizon **42**, 155
 South Pole Telescope 6
 spectral line 90
 spectroscopy **89**
 spherical harmonics **18**, 18–22
 standard deviation **93**
 standard error on the mean **93**
 star formation efficiency **153**
 Stefan–Boltzmann law 40
 stellar feedback **143**
 stellar mass **150**
 stellar mass function **150**, 150–154

Index

stellar merger 65
stellar remnant **145**
Strömgren sphere **137**
 radius 137
strong interaction 59
structure formation 107, 117, 149
super-horizon perturbations 32
supermassive black hole (SMBH) **146**
supernova 136, 145
 Type Ia 54
surface of last scattering **4**, 10
synchrotron radiation **13**, 146

Thomson cross-section 87
Thomson scattering 3, 33, 64, 86
tidal stream **141**
tracer 106, 136, 147, 151
tritium 77
two-point correlation function **154**
Type Ia supernova 54

unification 68

velocity dispersion **101**
virial theorem **100**, 132
virialisation 120
volume integral 100

weak interaction 59, 63
white dwarf 145
Wilkinson Microwave Anisotropy Probe
 (*WMAP*) 7, 8, 53
WMAP *see* *Wilkinson Microwave Anisotropy Probe*

Zeldovich pancake **121**