S385

Cosmology and the distant Universe

The Distant Universe

The Open University

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 2024.

Edited, designed and typeset by The Open University, using LaTeX.

Printed and bound in the United Kingdom by Hobbs the Printers Limited, Brunel Road, Totton, Hampshire SO40 3WX

ISBN 978 1 4730 3572 0

1.1

Contents

Introduction

In this book you will study a series of connected topics in the area of extragalactic astrophysics (the study of the Universe beyond the Milky Way). Each one involves exploring astrophysical environments in the context of galaxy evolution, and through consideration of the microphysics of radiation and matter, and their interaction. You will also revisit the topics of special and general relativity, and their impact on astronomical observations.

There are five chapters in *The Distant Universe*, as described below.

- Chapter 1 explores the epoch in the history of the Universe called cosmic dawn, during which the first stars and galaxies formed, and supermassive black holes grew together with their host galaxies.
- Chapter 2 describes the phenomenon of gravitational lensing, in which light from more-distant objects is bent in the vicinity of large masses, causing a variety of important observational effects.
- Chapter 3 provides an in-depth exploration of galaxy clusters, including how environmental factors affect galaxy evolution, and how galaxy-cluster observations are used to study dark matter.
- Chapter 4 considers the physics of the highly energetic jets that are ejected from supermassive black holes. You will read about their relativistic speeds and radiation processes (and investigate how these affect observations), and the importance of jets for galaxy evolution.
- Chapter 5 explores the explosive phenomenon of gamma-ray bursts, including their observational properties, and the best current models to explain their underlying physics and relation to star and galaxy evolution.

As with *Cosmology* Parts 1 and 2, the exercises in each chapter are an important element of your learning, with full solutions provided at the end of the book. The table of physical constants is also 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 *Cosmology* books, which you may find useful to revisit to remind yourself of the relevant underlying concepts. All other cross-references to content within chapters relate implicitly to this book, instead.

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.

Chapter 1 Cosmic dawn

Modern telescopes allow us to peer deep into the distant reaches of the Universe. The finite speed of light means that, by doing so, we are also observing galaxies as they appeared at much earlier times in the Universe's history. This chapter focuses on a period in the early Universe that is known as cosmic dawn, when the very first galaxies and black holes formed and ionised their surroundings. Remarkably, we now have the technology to find and study galaxies so distant that their light was emitted during the cosmic dawn era.

Objectives

Working through this chapter will enable you to:

- explain the significance of the period of reionisation for the evolution of the Universe, and summarise the main observational knowledge we have about when reionisation took place
- estimate the relative contributions of galaxies and quasars to reionisation in the early Universe
- describe the key methods used to find and study the earliest galaxies
- explain the significance of the Eddington limit for the growth of black holes in the early Universe
- summarise current observational and theoretical knowledge about the earliest black holes
- describe and investigate the key relationships between the properties of galaxies and their central black holes
- critically compare theories for the origin and growth of the first supermassive black holes (SMBHs).

1.1 A timeline of cosmic dawn

Figure 1.1 shows a timeline for how the baryonic gas in the Universe is thought to have evolved, focusing on the stages after recombination.

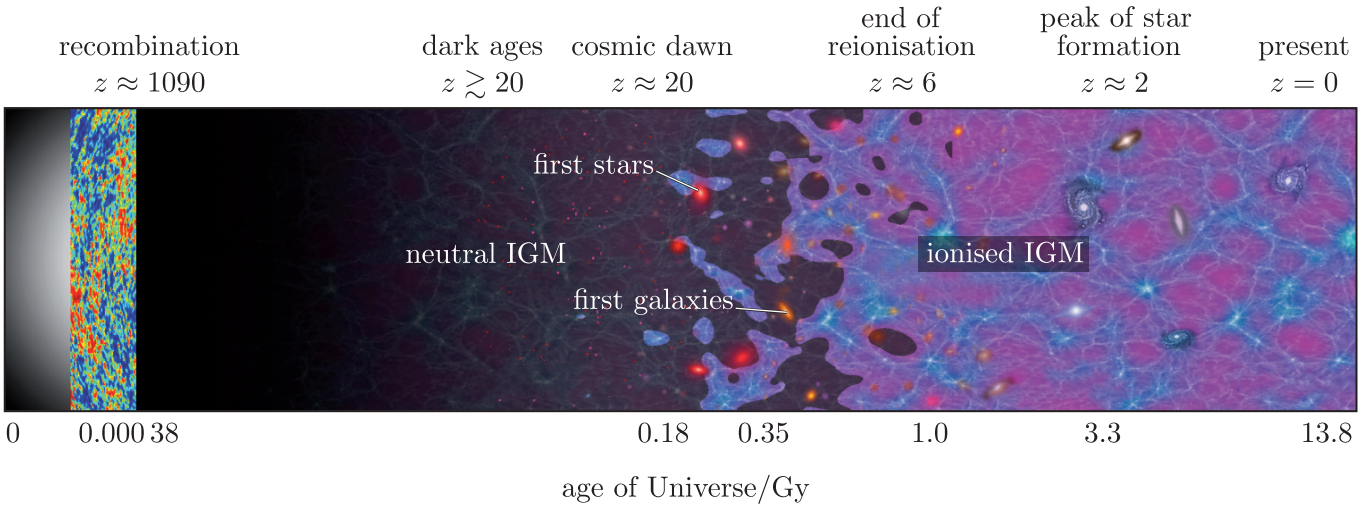


Figure 1.1 A timeline of how baryonic material evolved with cosmic time.

As you may recall from your study of *Cosmology*, the dark ages refers to the period in which matter clumped together under gravity prior to the production of the first stars.

- Why was the baryonic gas neutral during the dark ages?
- The gas became neutral at recombination, when electrons and ions came together to form atoms. The only photons present during the dark ages were the cosmic microwave background (CMB) photons, which were neither energetic nor numerous enough to ionise the gas.

Cosmic dawn describes the era during which 'the lights turned on': in other words, the period when the first stars and galaxies were formed and began to emit large amounts of radiation. The first stars (indicated in bright red in Figure 1.1) are thought to have formed when $z \lesssim 20$ and the Universe was $\sim$150–200 million years old; the first galaxies (shown as orange objects in the figure) appeared a little later. The purple regions, and the increasingly connected web-like structure from $z \approx 20$ onwards, represent an important phase change that affected the neutral intergalactic medium (IGM). Reionisation caused the gas to change gradually from being wholly neutral (in the dark ages) to fully ionised, and therefore transparent to optical light.

For simplicity in this chapter we consider the intergalactic gas to be comprised only of hydrogen, and neglect the smaller helium contribution. The photoionisation of hydrogen gas requires photons with an energy $E > 13.6\,\text{eV}$.

- Which parts of the electromagnetic spectrum correspond to this photon energy?

- A photon energy of 13.6 eV corresponds (via $E = hc/\lambda$) to a wavelength of $\lambda = 91.2$ nm. Ionising photons are therefore in the UV, X-ray and γ-ray parts of the spectrum.

In the remainder of this section we will explore how cosmic dawn and reionisation proceeded.

1.1.1 The first stars and galaxies

The stars we observe with our telescopes are usually either part of the Milky Way or of another galaxy. However, the very first (Population III) stars are thought to have formed *before* the first galaxies. There are two reasons for this. The first is that the gradual growth of dark-matter halos over cosmic time meant that the structures present at these early times had not yet reached the gas masses of present-day galaxies.

The second reason relates to the physics of star formation. In the present-day Universe, star formation proceeds via the fragmentation of gas clouds into smaller clumps, leading to clusters of stars each of which has an individual mass in the range $\sim$0.1–150 $M_\odot$. However, the fragmentation process is highly dependent on how the gas radiates away energy and cools.

- What properties of the gas affect its cooling rate?
- The cooling rate depends on gas density and temperature, but crucially also on metallicity, which controls the cooling mechanisms that can operate – see *Cosmology* Section 11.1.3).

It is the dependence on the presence of metals (elements other than hydrogen and helium) that differentiates the formation of the first stars and present-day ones. Population III stars formed from 'pristine' gas with a metallicity, $Z \approx 0$. Without heavier elements, the gas could only cool through less-efficient processes such as H_2 cooling, resulting in greatly reduced fragmentation.

The first generation of stars are expected to have typical masses of $\sim 10^3\,M_\odot$, which is far greater than even the most massive present-day stars. Their high mass made them both luminous and relatively short lived. These stars are thought to have played an especially important role in cosmic evolution.

Role of first stars in cosmic evolution

- The luminosity of the first stars reionised their surroundings and produced bubbles of relatively hot gas that were transparent to radiation.
- The energy released in supernovae at the end of these stars' short lives expelled surrounding gas from dark-matter halos. This expulsion of gas initially acted to slow down the gravitational collapse of surrounding gas to form further generations of stars, affecting the eventual properties of the first galaxies.

- The supernova explosions also enriched the surrounding gas, which enabled cooling via line emission from metals to take place. This meant more fragmentation of star-forming gas clouds could occur, so subsequent generations of stars had lower masses.
- The remnants of the first stars are expected to be intermediate-mass black holes, which are one possible seed population for the black holes that are observed in the centres of both nearby and distant galaxies (a topic discussed further at the end of this chapter).

The first galaxies started to form after the first generation of stars reached the end of their lives. As dark-matter halos grew in mass, the mean temperature of the baryonic gas within the halos increased in accordance with the virial theorem, leading to higher cooling rates. This produced reservoirs of cold, dense molecular gas that fuelled increasing star-formation rates. Figure 1.2 illustrates the changing star- and galaxy-formation processes across the period of reionisation. As more cooling processes were able to operate, increasingly high-mass systems could form.

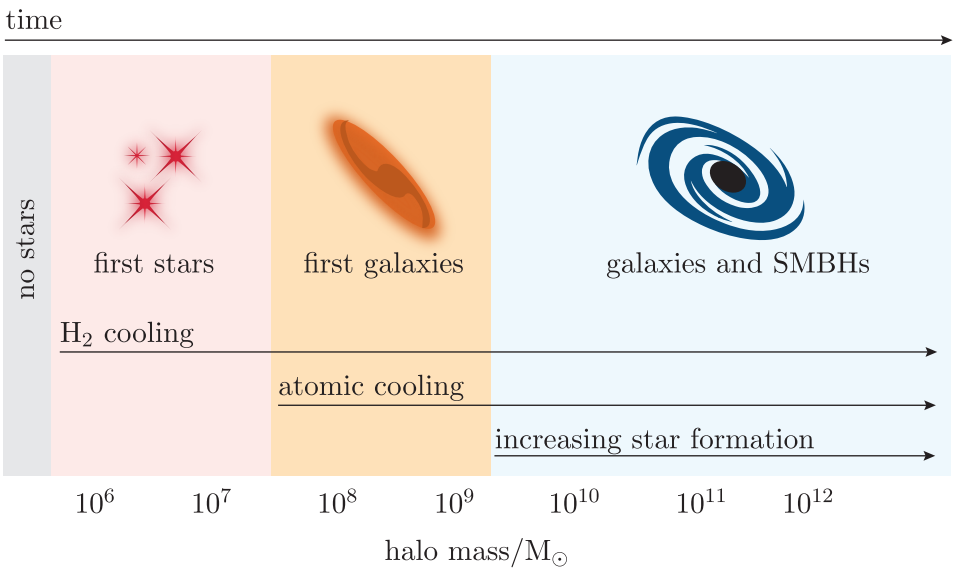


Figure 1.2 The changing processes of star and galaxy formation during cosmic dawn, with the maximum halo mass increasing with time.

1.1.2 Observing reionisation

One method of testing theories of cosmic dawn is to investigate the ionisation history of the intergalactic medium (IGM) in order to trace the influence of stars, galaxies and black holes over time. There are several powerful techniques to do this, including making inferences about the timing of reionisation from the CMB angular power spectrum and from radio measurements of redshifted atomic hydrogen signatures. However, in this section we will focus on the information that can be obtained from spectroscopic measurements of quasars.

Online resources: active galaxies

If you are unfamiliar with active galaxies and quasars, or would like a reminder, you may find it useful to look at the online module resources on this subject, which are taken from the Stage 2 astronomy curriculum.

Observations of quasars reveal that the IGM has been fully ionised for a large fraction of the time that is thought to have elapsed since the epoch of last scattering (when the CMB was produced). As the light from a distant quasar travels towards us it is absorbed at particular wavelengths when it encounters intervening neutral gas, in regions where small clumps of material have become dense enough to cool out of the ionised IGM. The absorption creates specific features in the resulting spectrum that provide information about the intervening gas. A particularly important series of atomic transitions in hydrogen gas is the Lyman series.

Lyman series

The Lyman series corresponds to transitions in atomic hydrogen between the $n = 1$ ground-state energy level and higher energy levels.

Lyman-α emission occurs when an electron decays from $n = 2$ to $n = 1$ and releases a photon with a (rest-frame) wavelength of $\lambda_{\text{em}} = 121.6\,\text{nm}$. Conversely, Lyman-α absorption occurs when a nearby photon is absorbed and excites the electron from $n = 1$ to $n = 2$. Features corresponding to both processes are commonly seen in quasar spectra.

Lyman-β transitions occur between $n = 3$ and $n = 1$, and correspond to a rest-frame wavelength of 102.6 nm.

Finally, as you read in *Cosmology* Chapter 11, the Lyman limit is at 91.2 nm, which is the wavelength needed to ionise a hydrogen atom.

- At what wavelength would you except to observe the Lyman-α emission line in the spectrum of a quasar at $z = 2$?
- Observed and emitted wavelengths are related to redshift via $\lambda_{\text{obs}} = \lambda_{\text{em}}(1 + z)$, so the line would be observed at $\lambda_{\text{obs}} = 364.8\,\text{nm}$.

Figure 1.3 illustrates the path of light from a quasar, and how intervening gas clouds can affect the spectrum (plotted as flux per unit wavelength, F_λ, versus λ_{obs}). In this scenario a Lyman-α emission line is produced from the emitting gas at the distance of the quasar, and is then redshifted as the light travels towards Earth. Lyman-α absorption then also occurs when the light passes through hydrogen gas clouds at particular intermediate distances, creating observed dips in flux at different wavelengths from the emission line (which was produced at the quasar itself).

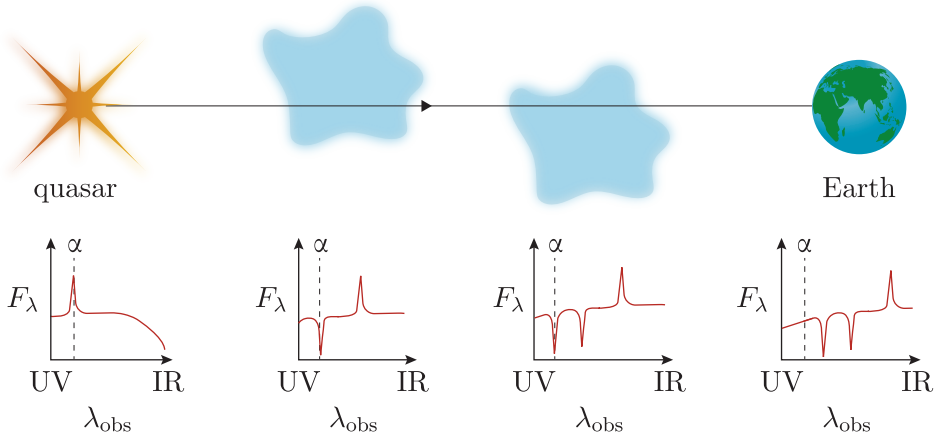


Figure 1.3 The path of light from a distant quasar to the Earth, resulting in a series of absorption features in the quasar spectrum from clouds at different distances. The plots show the spectrum as viewed at each location.

Each absorption feature will be redshifted by a different amount as the light travels to Earth, because of the different distances of intervening hydrogen clouds. This leads to a rich set of features in the quasar's spectrum known as the **Lyman-α forest**. Figure 1.4 shows an example quasar spectrum for a very high-redshift system, with the Lyman-α emission line identified. The many narrow lines to the left of this feature are the Lyman-α forest.

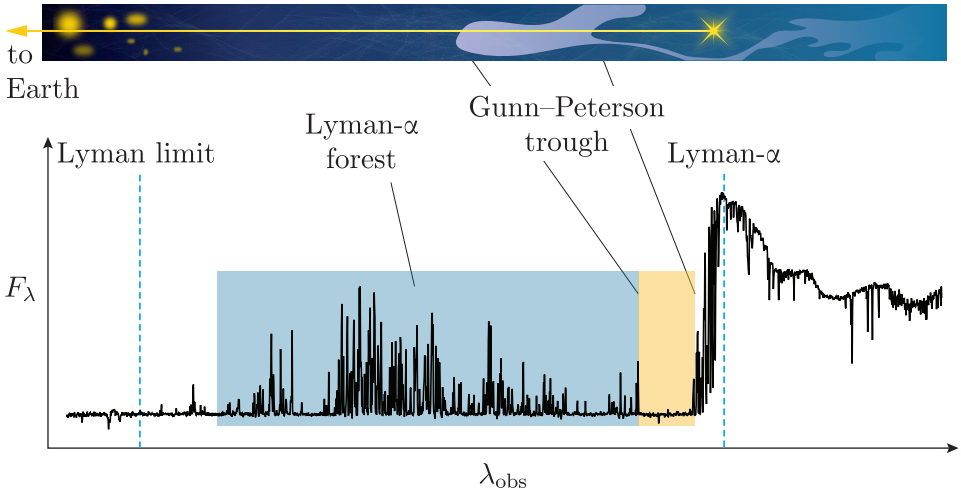


Figure 1.4 A typical high-redshift quasar spectrum showing Lyman series transitions and a Gunn–Peterson (G–P) trough, with the latter feature caused by the presence of neutral gas over a substantial redshift range in the IGM. The schematic above the spectrum shows the corresponding regions through which the light passed, with regions of neutral gas shown in paler blue/purple.

The most interesting feature for studying cosmic dawn in Figure 1.4 is the region of near-zero flux to the left of the Lyman-α line labelled as the **Gunn–Peterson trough**. This feature indicates that the light passed through an extended region of neutral (un-ionised) gas in the IGM at high redshifts. By comparing the wavelength range of the Gunn–Peterson (G–P) trough observed for quasars at different redshifts it is possible to determine when the period of reionisation ended and the intergalactic medium was fully ionised.

Figure 1.5 shows a series of high-redshift quasar spectra, presented in order of decreasing redshift. Those at $z > 6$ show a G–P trough to the left of their Lyman-α line, while those at lower redshifts show a forest of lines, to a greater or lesser extent, instead of a flat trough. These quasar observations therefore show that the IGM starts to contain a sizeable quantity of neutral gas above $z \approx 6$, which marks the end of the period of reionisation.

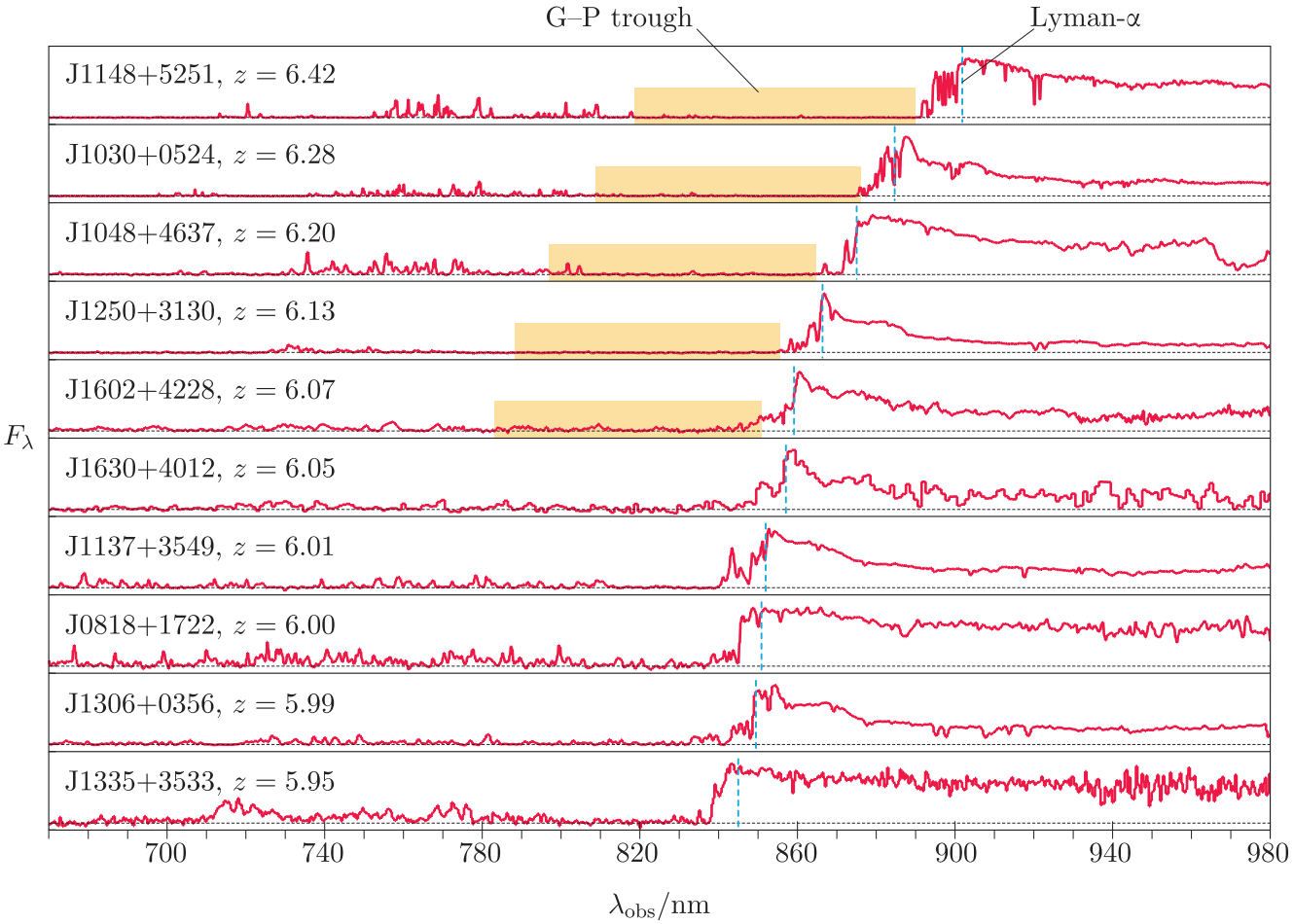


Figure 1.5 A sample of high-redshift quasar spectra, arranged in order of decreasing redshift. The G–P troughs and Lyman-α lines are also shown, where appropriate.

1.1.3 Sources of reionisation

The observational methods described in the previous section give us a firm idea that reionisation was complete by $z \approx 6$. The two candidates for reionising most of the gas in the Universe during this period are the growing stellar populations within galaxies, and quasars (which derive their luminosity from accretion of matter onto a central black hole). We will now consider the roles of these two sources of photons, following arguments presented by Ryden (2017).

We will consider the reionisation process in a *co-moving* volume at cosmic dawn. A co-moving volume of $1\,\mathrm{Mpc}^3$ refers to a region that was smaller at an earlier point in history, but will expand to a volume of $1\,\mathrm{Mpc}^3$ at the present day. (Recall the definition of co-moving distances from *Cosmology*, Chapters 3 and 5.) We can therefore assume that a typical such region contained the same *number* of baryons as it does in the present-day Universe, although the *number density* was historically higher, since the true physical volume was smaller at that earlier time.

At the present day, the co-moving volume has a baryon number density, $n_{\mathrm{b},0}$, of

$$n_{\mathrm{b},0} = \frac{\Omega_{\mathrm{b},0}\rho_{\mathrm{c},0}}{m_{\mathrm{p}}} \approx 7.4 \times 10^{66}\,\mathrm{Mpc}^{-3}$$

where $\Omega_{\mathrm{b},0}$ and $\rho_{\mathrm{c},0}$ are the present-day values of the baryon density parameter and critical density respectively, and m_{p} is the proton mass. Therefore a typical region of co-moving volume $1\ \mathrm{Mpc}^3$ will contain $\approx 7.4 \times 10^{66}$ baryons at cosmic dawn.

Next we can estimate how many ionising photons would be required to ionise such a region. The simplest assumption is that we just need one photon capable of ionisation for each baryon. However, we need to account for the fact that not all photons are able to escape from the galaxy in which they are produced – the photon **escape fraction**, f_{esc}, is an estimate of the proportion that do. The number of ionising photons, N_γ, needed to ionise a region of co-moving volume V_{comov} is therefore

$$N_\gamma = \frac{n_{\mathrm{b},0}V_{\mathrm{comov}}}{f_{\mathrm{esc}}} \tag{1.1}$$

The following example estimates the number density of massive stars needed to fully reionise a region of intergalactic gas.

Example 1.1

A co-moving volume of $1000\,\mathrm{Mpc}^3$ at $z = 8$ contains a number of galaxies, each of which is continually forming stars. The massive stars within the galaxies each produce ionising photons at a rate of $\dot{N}_\gamma = 5 \times 10^{48}\,\mathrm{s}^{-1}$. Answer the following questions, assuming that the escape fraction of photons from the galaxies is $f_{\mathrm{esc}} = 0.2$, that the stellar population in each galaxy is constant and that the stellar populations have been producing photons for 600 My.

(a) Estimate the number of massive stars that must be radiating within the volume in order to fully reionise the intergalactic gas.

(b) If galaxies at $z = 8$ each contain 20 000 massive stars (a similar number to the Milky Way), then what density of *galaxies* per co-moving cubic megaparsec would be needed to ionise all of the gas?

Solution

(a) The number of baryons to ionise in the intergalactic gas, N_b, is given by the present-day baryon density multiplied by the co-moving volume:

$$\begin{aligned} N_\mathrm{b} = n_{\mathrm{b},0} V_\mathrm{comov} &= 7.4 \times 10^{66}\,\mathrm{Mpc}^{-3} \times 1000\,\mathrm{Mpc}^3 \\ &= 7.4 \times 10^{69} \end{aligned}$$

Accounting for the escape fraction given, and using Equation 1.1, the number of ionising photons that need to be produced to ionise these baryons is

$$N_\gamma = \frac{N_\mathrm{b}}{f_\mathrm{esc}} = 3.7 \times 10^{70}$$

Individual massive stars produce ionising photons at a rate $\dot{N}_\gamma = 5 \times 10^{48}\,\mathrm{s}^{-1}$. Given that the stellar population in each galaxy is assumed to be constant, the total number of ionising photons produced is

$$N_\gamma = N_\mathrm{stars} \dot{N}_\gamma t$$

where N_stars is the number of massive stars within the volume and t is the period over which the photons are produced (in this case 600 My). Rearranging for N_stars and substituting in values for the other quantities, we find:

$$N_\mathrm{stars} = \frac{N_\gamma}{\dot{N}_\gamma t} \approx 390\,000$$

(b) In a co-moving volume of $1000\,\mathrm{Mpc}^3$ it would require $390\,000/20\,000 \approx 20$ present-day Milky-Way-like galaxies to fully reionise the interstellar gas. This is around 0.02 galaxies per co-moving cubic Mpc.

Now shifting our focus to quasars, the rate of ionising photon production for an individual galaxy of this type depends on its **bolometric luminosity** (the luminosity across all wavelengths) and the shape of its spectrum, which determines the fraction of the total number of photons that have energies in the UV and X-ray parts of the spectrum, i.e. those that are able to contribute to ionisation. This rate can be approximated for an individual quasar of luminosity L as:

$$\dot{N}_\gamma \approx 3 \times 10^{56}\,\mathrm{s}^{-1} \left(\frac{L}{10^{13}\,\mathrm{L}_\odot} \right) \tag{1.2}$$

It is thought that the escape fraction for quasar radiation may be considerably higher than for stars, perhaps as high as $f_{\rm esc} \sim 1$ for the most luminous quasars. In the following exercise you can make a rough estimate of the number density of quasars needed for reionisation.

Exercise 1.1

(a) Given a typical quasar bolometric luminosity of $L = 10^{39}\,\mathrm{W}$ and photon escape fraction $f_{\rm esc} = 1$, calculate the number of luminous quasars, $N_{\rm Q}$, needed to reionise a co-moving volume of $1000\,\mathrm{Mpc}^3$. Assume that quasars produce photons over a time period of 600 My and that the number of quasars remains constant.

(b) Hence determine the co-moving number density of quasars needed for quasars to reionise a $1000\,\mathrm{Mpc}^3$ co-moving volume (neglecting any contributions from the massive stars).

Exercise 1.1 and Example 1.1 suggest that the density of quasars needed to carry out reionisation alone is much smaller than for galaxies. However, luminous quasars are much rarer than ordinary galaxies, and so it is now thought that massive stars (and not quasars) are responsible for most of the reionisation at cosmic dawn. In the remainder of the chapter you will learn about direct observations of these early galaxies and black holes.

1.2 Finding the earliest galaxies

Telescope technology has now advanced to the point where we can *directly* observe galaxies at distances so large that we are seeing them at around the time of cosmic dawn, when it is thought that the first galaxies formed.

1.2.1 High-redshift galaxy surveys

In order to study the earliest galaxies it is necessary to first detect them in our images and then to distinguish them from the large numbers of similarly faint, but closer and less-luminous galaxies, which are found in the most sensitive (deepest) observations. Hence these studies require powerful telescopes. Figure 1.6 highlights a small number of very high-redshift galaxies found within a *JWST* deep-field survey image.

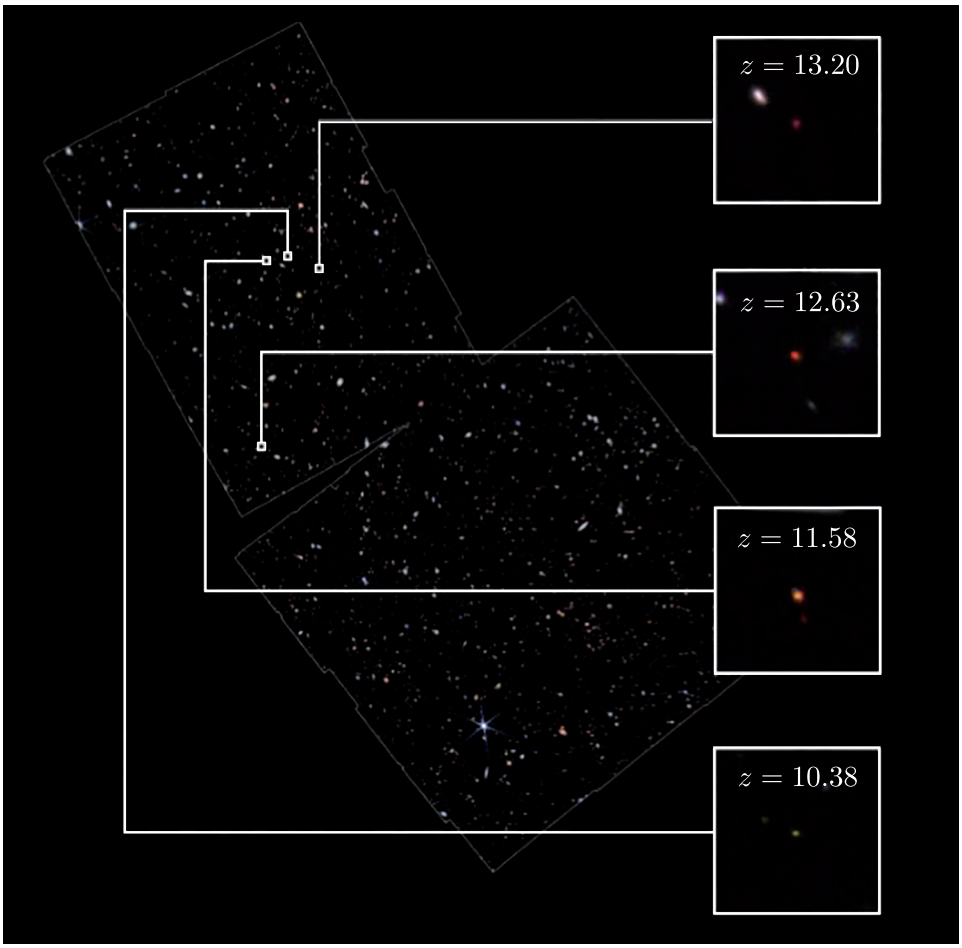


Figure 1.6 Examples of very high-redshift galaxies identified in deep *JWST* observations and confirmed with spectroscopic measurements. In each of the right-hand boxes the distant galaxy is the object at the centre of the image.

The only way to be certain of a galaxy's distance is to measure its redshift precisely from a spectrum, but this requires long observations and so cannot be done for all of the hundreds of faint galaxies in a deep-field image. However, **photometric redshifts** can be estimated based on measurements of galaxy colours (i.e. comparing their brightness through different telescope filters).

Figure 1.7 shows a histogram of these photometric redshift estimates for galaxies from the same survey as shown in Figure 1.6, and demonstrates that only a small number of the galaxies have redshifts that could place them well into the cosmic dawn era. The space density of galaxies beyond $z \approx 8$ is not yet well determined, but *JWST* is expected to make major progress in this area.

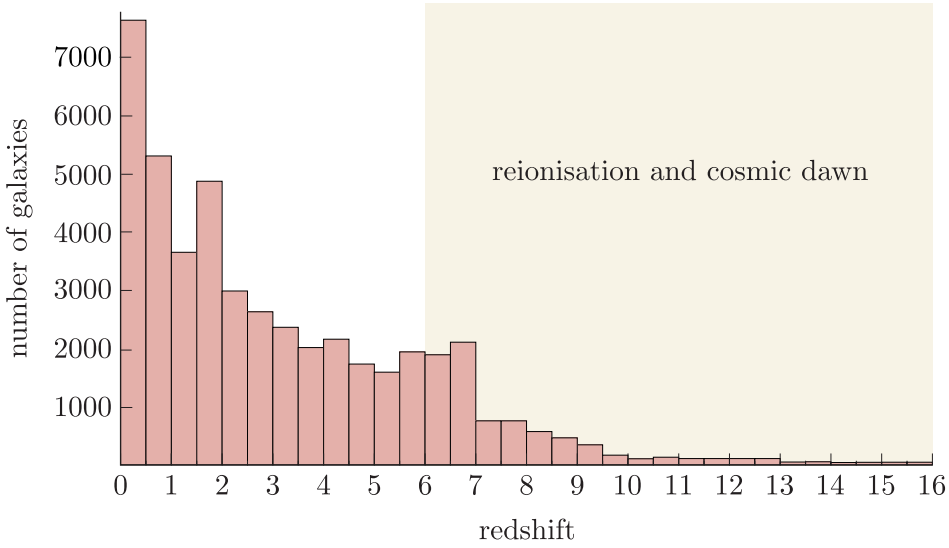


Figure 1.7 The distribution of estimated redshifts and lookback times for a typical deep field observed by *JWST*.

The next section explains in more detail how the colours of galaxies are used to narrow down the possibilities when trying to find the relatively rare examples that are at the highest redshifts.

1.2.2 The Lyman-break method

The primary method used to identify candidate distant galaxies is known as the **Lyman-break method** (or the dropout method). This is a selection method applied to deep surveys, typically of small areas of sky in directions away from the plane of the Milky Way.

The physics behind the Lyman-break method is related to the properties of Lyman series emission and absorption transitions. The **Lyman break** is a prominent feature in the spectra of typical star-forming galaxies. It is caused by absorption of the galaxy's radiation by gas, both within that same galaxy and in intervening regions at lower redshifts.

At photon energies *below* the Lyman limit (i.e. wavelengths longer than $\lambda_{\rm em} = 91.2\,{\rm nm}$) only specific wavelengths of light are absorbed. These interactions correspond to the Lyman-α, -β and higher transitions, and create 'dips' in a galaxy spectrum. Above this limit (i.e. at shorter wavelengths) photons will ionise the gas and so all wavelengths can be absorbed, which leads to a large drop in the galaxy flux that escapes.

Figure 1.8 shows the spectrum of a distant galaxy, with a sharp drop in flux at a characteristic wavelength corresponding to the Lyman break. All galaxy spectra contain this prominent feature, which will be redshifted according to the galaxy's distance from Earth. The Lyman-break method involves searching for galaxies at a particular redshift by comparing images made using telescope filters of different wavelengths.

If a galaxy is observed with a filter corresponding to a wavelength range *above* the (redshifted) location of the Lyman break then it will be bright, as is the case for the the image shown above the 'filter 2' range in Figure 1.8. If the same galaxy is observed with a filter *below* the break then it will be very much fainter or absent, as you can see in the image corresponding to the 'filter 1' range. The aim of the method is to find galaxies whose Lyman break is close to the boundary between the filters.

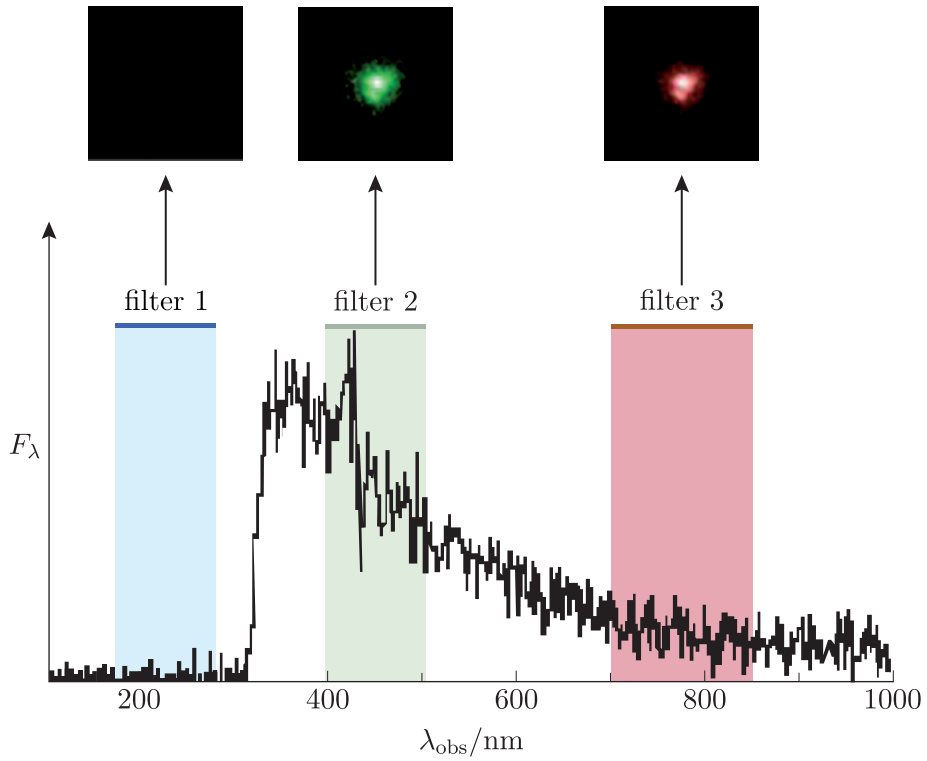


Figure 1.8 Images of a typical star-forming galaxy using three telescope filters, compared with a full galaxy spectrum. Comparing the brightness of the three images allows a rough estimate of the galaxy's redshift because of the observed absence of emission in the image from the lowest-wavelength filter.

- If the sharp drop in flux in Figure 1.8 corresponds to the Lyman break, what is the approximate redshift of this galaxy?

- The observed wavelength of the break is $\lambda_{obs} \approx 300$–$350\,\text{nm}$. Using $\lambda_{obs} = \lambda_{em}(1+z)$ with $\lambda_{em} = 91.2\,\text{nm}$, the galaxy redshift is $z \approx 2.3$–2.8.

As we consider objects at increasingly high redshift, the Lyman break shifts towards longer wavelengths and it becomes necessary to observe in the near-infrared part of the spectrum. *JWST* was designed to enable this method to be applied to find galaxies with much higher redshifts.

- Would you expect the G–P trough to be relevant for galaxies as well as quasars and, if so, how would this affect the Lyman-break method?

- Yes, for galaxies at redshifts when the Universe was not fully ionised, i.e. $z \approx 6$, absorption due to intervening neutral gas (which produces the G–P trough for quasars) should still occur. This would move the rest-frame wavelength of the observed break in the spectrum from 91.2 nm (the Lyman break) to 121.6 nm.

Figure 1.9 shows the spectra of two very high-redshift galaxies measured by the *JWST* NIRSpec instrument, which were first identified via the dropout method. Due to the galaxies' high redshifts, the spectral break occurs at the Lyman-α redshift (corresponding to a wavelength of 121.6 nm) and not at the Lyman limit, as explained above.

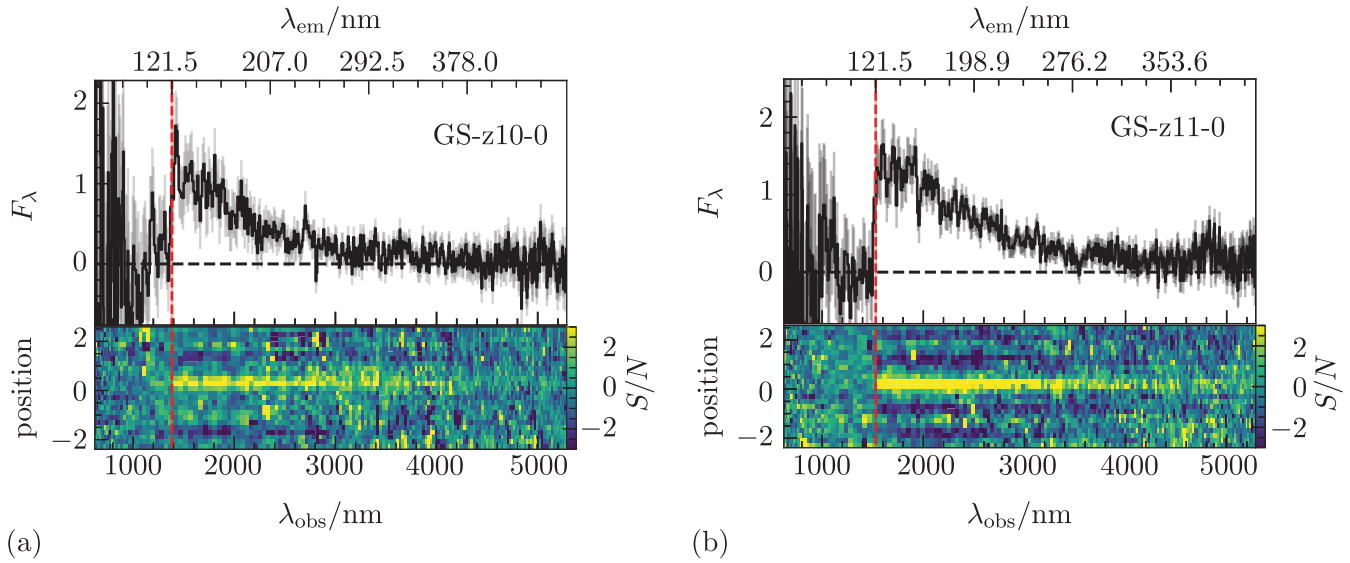


Figure 1.9 The spectra of two $z > 10$ galaxies: (a) GS-z10-0 and (b) GS-z11-0. The lower panels show the signal-to-noise ratio (S/N) of the detected emission against position, in a slice across the source. There is a sharp drop in S/N to the left of the rest-frame Lyman-α wavelength (marked by vertical red lines in the top panels).

The sharp breaks observed between wavelengths of 1 and 2 μm in both cases correspond to absorption at wavelengths shorter than the rest-frame Lyman-α line. By comparing the rest-frame (emitted) wavelengths with the observed wavelengths, the redshifts of GS-z10-0 and GS-z11-0 are confirmed to be $z \approx 10.4$ and $z \approx 11.6$, respectively. The light we're measuring from these galaxies was produced when the Universe was only ∼400 million years old. *JWST* is being used to search for galaxies at redshifts even greater than these two examples.

1.2.3 Lensing effects on distant galaxies

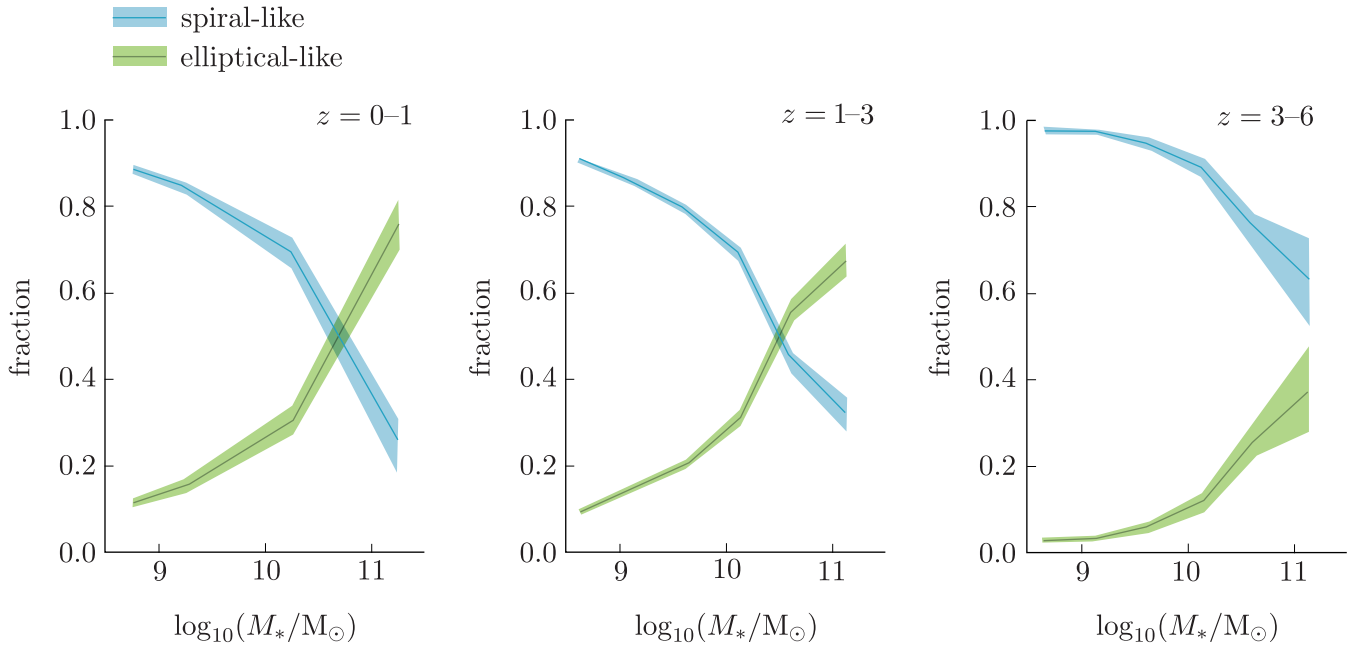
Another powerful method for finding and studying the most distant galaxies relies on **gravitational lensing**, which is the amplification of light from a distant object along a path that passes close to a nearer massive object. The amplified brightness of lensed galaxies means that we are able to observe fainter, more distant systems whose unamplified light would be undetectable.

Some of the highest-redshift galaxies so far detected by *JWST* are known to be lensed. The Lyman-break method can still be applied to such galaxies because lensing affects all wavelengths equally, and so does not alter a galaxy's spectrum.

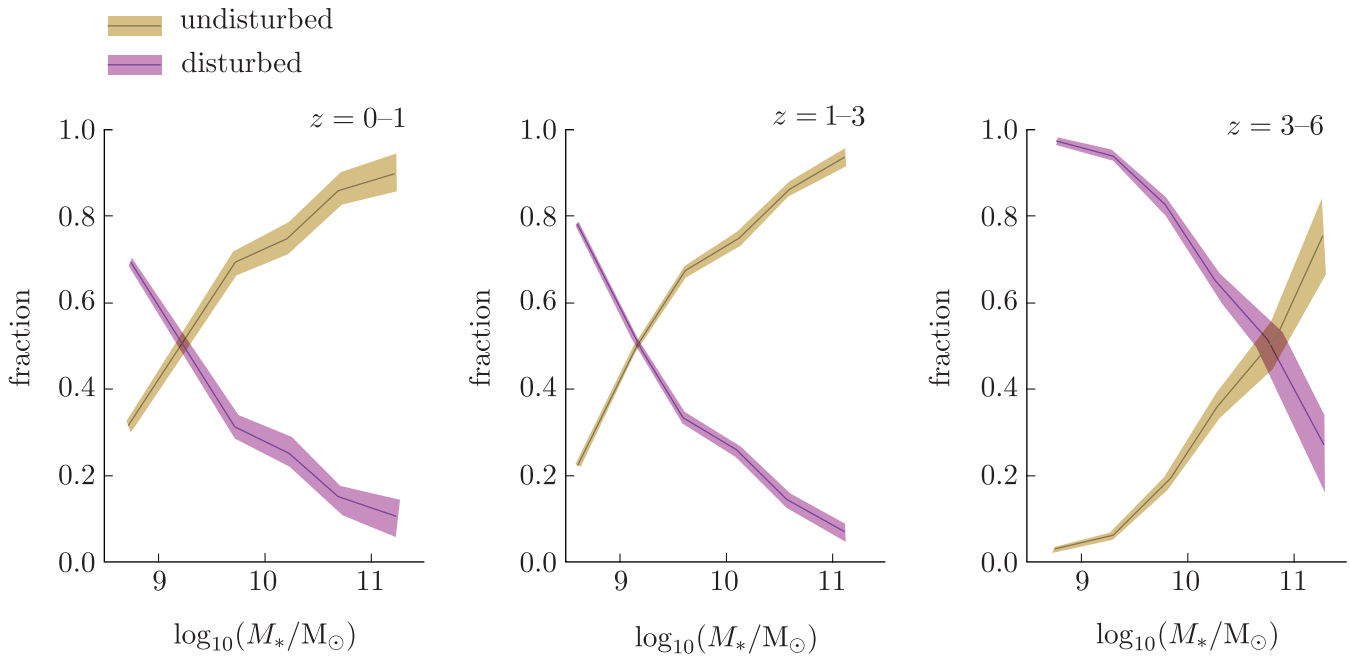
The effects of lensing are sometimes obvious. For example, lensing can lead to galaxy images that are duplicated or highly distorted (see the next chapter for more information). However, sometimes the effects of lensing are subtle, and if not identified or corrected for they can lead to bias in estimates of galaxy properties or luminosity functions at different redshifts.

1.2.4 Properties of the highest-redshift galaxies

Measurements of how redshift influences the properties of galaxies (including colours, galaxy luminosity functions, metallicities, star-formation rates and stellar population measures) have been used to develop modern theories of galaxy evolution, and to test numerical simulations. Figure 1.10 shows the results of one study, which examines how the fractions of different types of galaxies evolve with redshift and galaxy mass. Panel (a) differentiates between spiral-like and elliptical-like galaxies across three redshift ranges, whereas panel (b) compares galaxies on the basis of whether or not their morphology has been disturbed by interactions with other galaxies, such as mergers. In each case the horizontal axis shows the stellar mass of the galaxies.



(a)



(b)

Figure 1.10 The evolution of galaxy structure with redshift and galaxy mass, as measured by *JWST*. The redshift ranges for the samples increase from left to right across the three columns. Panel (a) compares disc-dominated (e.g. spiral-like) and bulge-dominated (e.g. elliptical-like) galaxies, and panel (b) compares galaxies with disturbed and undisturbed morphologies.

- Over what redshift range does the galaxy structure appear to change most dramatically in Figure 1.10, and in what way?

- For both panels (a) and (b) there is relatively little difference between the proportion of different galaxy types in the $z = 0$–1 and $z = 1$–3 plots, but a much larger difference in these proportions is seen in the final $z = 3$–6 plot. The fraction of spiral-like galaxies becomes very high at all masses at the highest redshifts, where the fraction of disturbed galaxies dominates over a wider range of galaxy masses.

Overall the results of these types of galaxy morphological study are in good agreement with models of galaxy evolution, in which galaxy mergers occur frequently at higher redshifts, and spiral structure disappears (preferentially in the most massive systems) as galaxies grow via mergers.

Galaxy evolution models also predict that galaxy luminosity functions (see *Cosmology* Chapter 11) will change significantly with time. Measuring the galaxy luminosity function provides a census of the population of galaxies at each redshift. Figure 1.11 shows the evolution of the UV-luminosity function of galaxies, out to the highest redshifts for which reliable information is available to date. Note that these are plotted as a function of magnitude, so that the lowest luminosities are to the right in the plot.

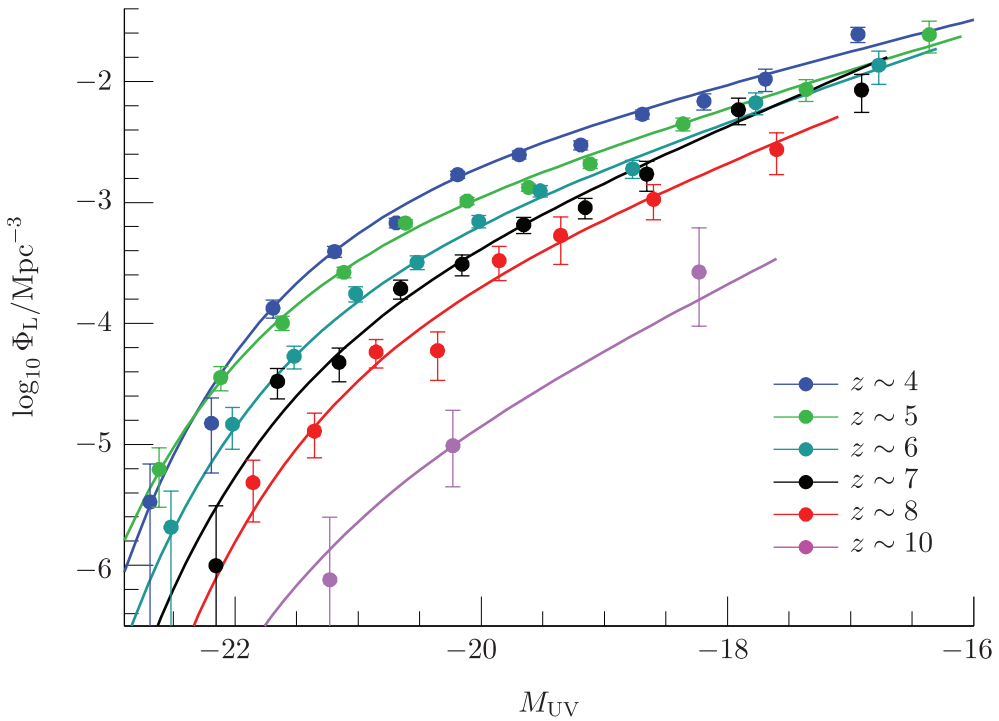


Figure 1.11 UV-luminosity functions for high-redshift galaxies, showing evolution of the overall number density.

It is evident that the shape and the normalisation of the UV-luminosity function – namely the overall scaling in the vertical direction – changes with time. The number density of galaxies (within the range it has been possible to measure) is much lower at $z \sim 8$ than at $z \sim 4$, but also the relative proportions of brighter and fainter galaxies has changed. The shape of the function is flatter at lower redshift, so that there is a higher proportion of more luminous galaxies (i.e. to the left of the plot) relative to fainter ones. This is as expected if the most massive and luminous galaxies were formed gradually over many billions of years via mergers.

In Example 1.1 we estimated that a galaxy number density of about $0.02\,\mathrm{Mpc}^{-3}$ is needed to reionise the Universe at $z = 8$. The luminosity function plotted in Figure 1.11 at $z \sim 8$ shows that the number density of galaxies starts to reach this value at the very right-hand side of the plot, corresponding to the lowest-luminosity galaxies (much less luminous than the Milky-Way-like galaxy we used for the reionisation estimate in the earlier example).

This might suggest it would have been difficult for galaxies at $z = 8$ to produce enough radiation to reionise the gas. However, more detailed reionisation calculations indicate that the galaxy population at $z \sim 8$ *would* be sufficient, provided the observed steep slope of the luminosity function continues to fainter luminosities than are shown in Figure 1.11 (because they are too faint to have been observed at this distance). It is a key goal of ongoing *JWST* research to extend the current measurements of galaxy luminosity functions out to even higher redshifts, to the very earliest galaxies.

1.3 The earliest black holes

Observations of quasars and high-redshift galaxies show that supermassive black holes (SMBHs) have been present since the early Universe. This raises the interesting question of when and how the SMBHs in galaxy centres formed. In this section we will examine correlations between the properties of SMBHs and their host galaxies, and what these connections tell us about how both parties evolve. We will also explore theories of how the first such black holes formed.

1.3.1 Black hole activity across cosmic time

The evolution of black holes can be investigated by studying quasars and active galactic nuclei (AGN) at optical, infrared, X-ray and radio wavelengths. Although we now believe that all galaxies have a central black hole, it is in AGN that we see the most direct evidence for their presence, via the large amounts of radiation associated with the infall of gas onto the black hole.

Figure 1.12 shows how the number density of quasars, $n_{\rm Q}$ (selected below a certain UV absolute magnitude, $M_{\rm UV} < -23$) evolves with redshift, based on careful measurements of the quasar luminosity function at different redshifts. The black diagonal line is a power-law model fitted to the high-redshift data, and has the form:

$$n_{\rm Q} \propto 10^{-0.78z} \tag{1.3}$$

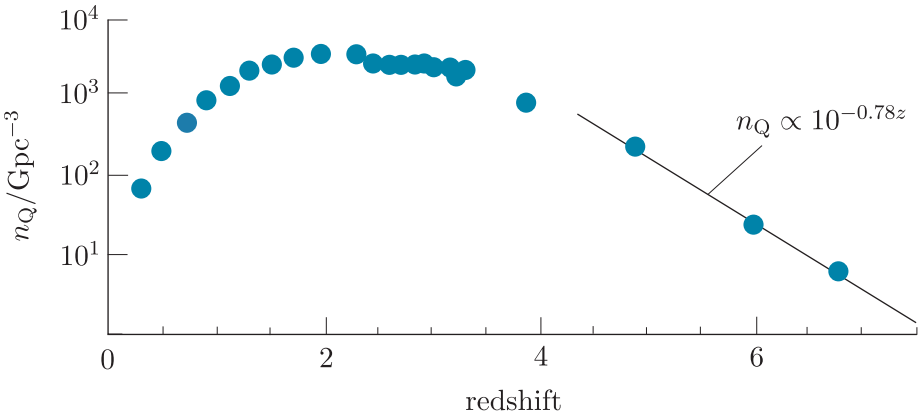


Figure 1.12 The evolution of the number density of quasars of $M_{\rm UV} < -23$ with redshift.

Exercise 1.2

Use Figure 1.12 and Equation 1.3 to answer the following questions.

(a) Summarise, in a few sentences, how quasar number density changes with redshift, and comment on how the present-day density of quasars compares with the density at the end of cosmic dawn.

(b) Estimate the value of $n_{\rm Q}$ at $z = 8$.

(c) Compare your answer to part (b) with the results of Exercise 1.1 (namely that $n_{\rm Q} \approx 5 \times 10^{-6}\,{\rm Mpc}^{-3}$ is needed for reionisation at $z = 8$), and comment on whether the observed quasar population evolution suggests they played an important role in reionisation.

The evolution of the quasar population is one piece of information about how black holes change over cosmic time. However, to study how black holes evolve in galaxies we need methods to estimate their masses, rather than just their (easier-to-measure) luminosities.

1.3.2 Measuring black-hole masses

One simple way to estimate the masses of black holes is to assume that quasars are radiating at the **Eddington limit** – the maximum luminosity that can be radiated by an accreting black hole of a particular mass. This limit is reached when the outward radiation pressure from the luminous source balances the gravitational inflow of mass that fuels the black hole.

The main source of an active galaxy's radiation is the conversion of a portion of the gravitational potential energy released by the gas that is falling into a central black hole. The change in gravitational potential energy $\Delta E_{\rm acc}$ of a parcel of gas of mass m, accreting from infinity to the event horizon is:

$$\Delta E_{\rm acc} = \frac{GM_{\rm BH}m}{R_{\rm S}} = \frac{1}{2}mc^2 \tag{1.4}$$

where $M_{\rm BH}$ is the black-hole mass and $R_{\rm S} = 2GM_{\rm BH}/c^2$ is its Schwarzschild radius.

In order words, the release of gravitational potential energy *could* make available up to half of the rest-mass energy of the accreted gas. By comparison, the release of rest-mass energy by nuclear fusion from hydrogen to helium (e.g. as in the core of Sun) liberates only $\Delta E = 0.007mc^2$, so less than 1% of the rest-mass energy of the gas undergoing fusion. Accretion is therefore a very powerful energy source!

The **accretion rate** is defined as the time derivative of infalling mass, $\dot{m} = {\rm d}m/{\rm d}t$. Since the available energy depends only on the accreting mass, the accretion rate determines the AGN luminosity L, so that

$$L = \frac{\Delta E}{\Delta t} = \eta \dot{m} c^2 \tag{1.5}$$

where η is the efficiency of the conversion of infalling mass into radiation, which is typically assumed be $\sim$0.1 (i.e. 10% of the infalling mass is converted to radiation, with the remainder accreted by the black hole).

The outward-acting force, $F_{\rm rad}$, due to the radiation pressure on a region of infalling mass at a distance R from the radiation source is

$$F_{\rm rad} = \frac{L\sigma_{\rm T}m}{4\pi R^2 c m_{\rm p}} \tag{1.6}$$

Here $\sigma_{\rm T}$ is the Thomson cross-section, and the gas is assumed to be pure hydrogen so that $m/m_{\rm p}$ provides an estimate of the number of electrons undergoing scattering by the radiation.

We can now equate this outward force with the inward gravitational force acting on the mass to determine the conditions under which the forces balance, and the accretion rate cannot increase:

$$\frac{L\sigma_{\rm T}m}{4\pi R^2 c m_{\rm p}} = \frac{GM_{\rm BH}m}{R^2}$$

Rearranging for L leads to the definition of the Eddington luminosity, which cannot be exceeded by a spherically symmetric accreting system.

Eddington luminosity

$$L_{\mathrm{Edd}} = \frac{4\pi G M_{\mathrm{BH}} m_{\mathrm{p}} c}{\sigma_{\mathrm{T}}} \tag{1.7}$$

The only non-constant parameter in Equation 1.7 is the black-hole mass, M_{BH}. The relation therefore provides a simple way to obtain a rough estimate of the black-hole mass of a quasar if it is assumed to be radiating at (or near to) the Eddington luminosity. This is a commonly used assumption, but it is important to bear in mind that not all active galaxies will be accreting at this rate, which means that mass estimates made this way can have a large uncertainty.

A more direct method of estimating black-hole masses uses spectroscopy to measure the width of emission lines originating in an AGN's **broad line region** – an area of gas clouds located close to the black hole. The linewidths are caused by Doppler broadening due to the rapid orbits of the gas clouds, and so the virial theorem can be used to relate the velocities measured from the emission lines to the central mass controlling the cloud motions.

This method requires more information than the Eddington luminosity calculation, but does not rely on the assumption that the black hole is accreting at the maximum rate. Indeed, when black-hole mass is measured it this way, it is then possible to compare the observed AGN luminosity with the Eddington luminosity to estimate the mass accretion rate.

1.3.3 The galaxy–black-hole connection

It is possible to make accurate measurements of black-hole masses in the local Universe. This allows us to establish tight correlations between the mass of a central black hole, M_{BH}, and galaxy properties such as the near-infrared (K-band) luminosity of the galaxy bulge, L_{K}, and velocity dispersion, σ_{v}. (As stated in *Cosmology* Chapter 9, σ_{v} is the typical speed in the radial direction of a group of objects, such as the galaxies in a cluster or, in this case, the stars in a galaxy.) Figure 1.13 shows the relationships between M_{BH} and the other two properties.

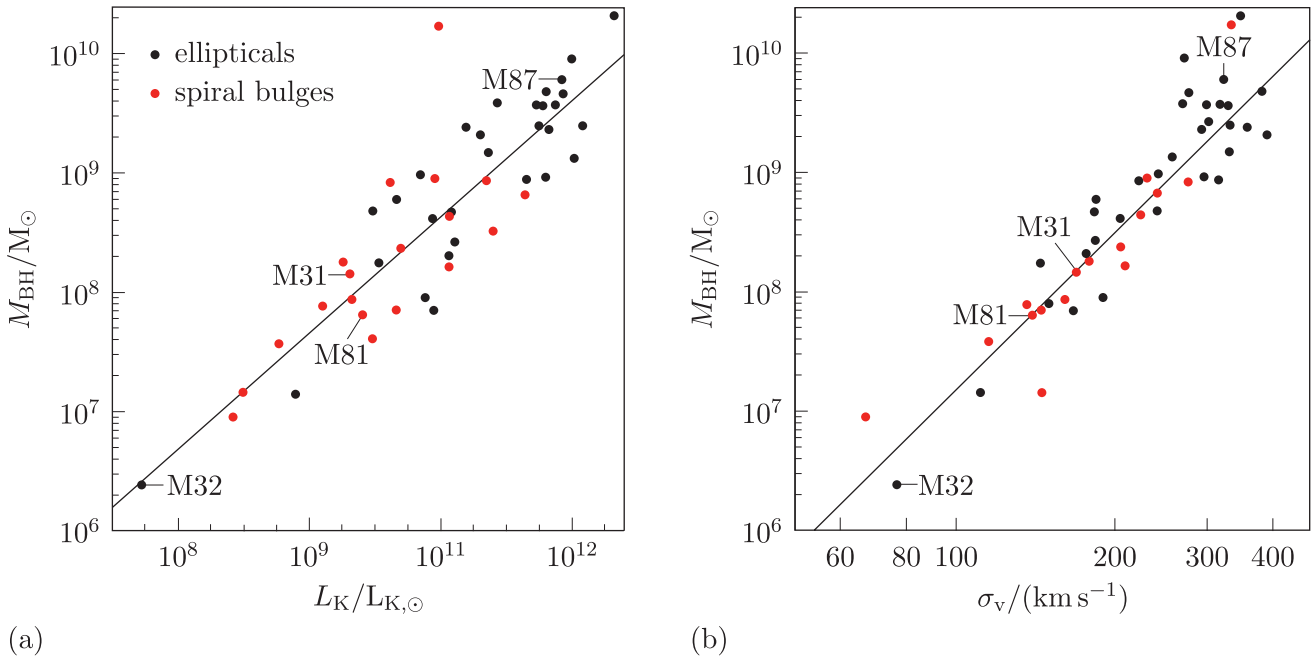


Figure 1.13 The tight relationship between black-hole mass and (a) K-band luminosity, and (b) velocity dispersion. Each point is an individual galaxy and is categorised by its morphology. Some better-known galaxies have been labelled in each plot.

The best-fitting relationships for M_{BH}–L_{K} and M_{BH}–σ_{v} (as plotted in the figure) are:

$$\frac{M_{\mathrm{BH}}}{10^9\,\mathrm{M}_\odot} = 0.54\left(\frac{L_{\mathrm{K}}}{10^{11}L_{\mathrm{K},\odot}}\right)^{1.2} \tag{1.8}$$

$$\frac{M_{\mathrm{BH}}}{10^9\,\mathrm{M}_\odot} = 0.31\left(\frac{\sigma_{\mathrm{v}}}{200\,\mathrm{km\,s^{-1}}}\right)^{4.4} \tag{1.9}$$

Here, $\mathrm{L}_{\mathrm{K}\odot}$ is the K-band luminosity of the Sun, and masses and luminosities are in solar units.

A further important relationship is observed between the mass of the black hole and the mass of the galaxy bulge (the **bulge mass**), where the latter encompasses the entire galaxy for ellipticals and the central region (excluding the spiral arms) for a spiral galaxy. The relationship is expressed as

$$\frac{M_{\mathrm{BH}}}{10^9\,\mathrm{M}_\odot} = 0.49\left(\frac{M_{\mathrm{bulge}}}{10^{11}\,\mathrm{M}_\odot}\right)^{1.2} \tag{1.10}$$

The following example explores how these relationships can be applied in order to understand whether a particular galaxy and its black hole are typical or unusual in terms of their relative masses.

Example 1.2

A galaxy has a measured K-band luminosity $L_{\rm K} = (8.5 \pm 0.6) \times 10^9\,{\rm L}_{{\rm K},\odot}$, a velocity dispersion $\sigma_{\rm v} = 110 \pm 5\,{\rm km\,s^{-1}}$, and a bulge mass $M_{\rm bulge} = (7.1 \pm 0.5) \times 10^9\,{\rm M}_\odot$.

(a) Use Equations 1.8 and 1.9 to make two estimates of the central black-hole mass for this galaxy, along with their associated errors. Comment on whether these estimates give consistent results.

(*Hint*: if a function $f \propto x^b$ then the fractional uncertainty in the function is $\Delta f/f = b\Delta x/x$.)

(b) Take the average of the two black-hole mass estimates from part (a) and calculate the associated error.

(*Hint*: if $f = (x + y)/2$ then $\Delta f = \sqrt{(0.5\Delta x)^2 + (0.5\Delta y)^2}$.)

(c) Using the result from part (b), predict the bulge mass for this galaxy. Comment on whether this is consistent (within the uncertainties) with the estimate for $M_{\rm bulge}$ provided in the question.

Solution

(a) Substituting in the provided value of $L_{\rm K}$ into Equation 1.8 gives a black-hole mass estimate of:

$$M_{\rm BH} = 0.54 \left(\frac{8.5 \times 10^9\,{\rm L}_{{\rm K},\odot}}{10^{11}\,{\rm L}_{{\rm K},\odot}} \right)^{1.2} \times 10^9\,{\rm M}_\odot = 2.80 \times 10^7\,{\rm M}_\odot$$

The luminosity has a fractional uncertainty of $\Delta L_{\rm K}/L_{\rm K} \approx 0.071$, so:

$$\Delta M_{\rm BH} = 1.2 \times 0.071 \times 2.80 \times 10^7\,{\rm M}_\odot = 0.20 \times 10^7\,{\rm M}_\odot$$

Similarly, the second estimate of black-hole mass comes from substituting in the provided value of $\sigma_{\rm v}$ into Equation 1.9:

$$M_{\rm BH} = 0.31 \left(\frac{110\,{\rm km\,s^{-1}}}{200\,{\rm km\,s^{-1}}} \right)^{4.4} \times 10^9\,{\rm M}_\odot = 2.23 \times 10^7\,{\rm M}_\odot$$

The velocity dispersion has fractional uncertainty of $\Delta\sigma_{\rm v}/\sigma_{\rm v} = 0.045$, so in this case

$$\Delta M_{\rm BH} = 4.4 \times 0.045 \times 2.23 \times 10^7\,{\rm M}_\odot = 0.45 \times 10^7\,{\rm M}_\odot$$

Hence the two estimates are consistent within the experimental error (the maximum acceptable value from the $\sigma_{\rm v}$ calculation agrees with the minimum acceptable value from the $L_{\rm K}$ calculation).

(b) The average black-hole mass from the two estimates is

$$\langle M_{\rm BH} \rangle = 2.5 \times 10^7\,{\rm M}_\odot$$

Using the hint provided, the associated error is

$$\Delta\langle M_{\rm BH} \rangle = \sqrt{(0.5 \times 0.2)^2 + (0.5 \times 0.4)^2} \times 10^7\,{\rm M}_\odot \approx 0.2 \times 10^7\,{\rm M}_\odot$$

(c) Rearranging Equation 1.10 for $M_{\rm bulge}$ and using $M_{\rm BH} = \langle M_{\rm BH} \rangle$ from part (b) gives:

$$M_{\rm bulge} = \left(\frac{2.5 \times 10^7\,{\rm M}_\odot}{0.49 \times 10^9\,{\rm M}_\odot} \right)^{1/1.2} \times 10^{11}\,{\rm M}_\odot = 8.4 \times 10^9\,{\rm M}_\odot$$

Now using the same error propagation formula as in part (a), the error on the bulge mass estimate, $\Delta M_{\rm bulge}/M_{\rm bulge}$, is given by:

$$\frac{\Delta M_{\rm bulge}}{M_{\rm bulge}} = (1/1.2) \times \frac{\Delta M_{\rm BH}}{M_{\rm BH}} = (1/1.2) \times \frac{0.2 \times 10^7\,{\rm M}_\odot}{2.5 \times 10^7\,{\rm M}_\odot} \approx 0.1$$

The lowest value within our estimated bulge mass range is therefore $M_{\rm bulge} - \Delta M_{\rm bulge} = 8.4 \times 10^9\,{\rm M}_\odot - 0.1 \times (8.4 \times 10^9\,{\rm M}_\odot) = 7.6 \times 10^9\,{\rm M}_\odot$. The highest value within the uncertainty range of the observational measurement of $M_{\rm bulge}$ given in the question is $M_{\rm bulge} + \Delta M_{\rm bulge} = 7.6 \times 10^9\,{\rm M}_\odot$.

We have therefore shown (with some effort!) that the black-hole mass estimates from part (a) imply a bulge mass that is consistent with the value given in the question, to within the quoted uncertainties.

Example 1.2 considered in some detail how different measurements relating to galaxy and black-hole properties can be compared. Try the following exercise to practise some similar calculations.

Exercise 1.3

Table 1.1 provides the measured bulge and black-hole masses of three nearby galaxies: A, B and C. For each galaxy, use the measured bulge mass to *predict* the black-hole mass (and its uncertainty) using Equation 1.10. Hence determine which of the galaxies have properties that are consistent with the relation and which appear to deviate from it.

Table 1.1 A set of observed galaxy bulge and black-hole masses.

Galaxy	measured $M_{\rm bulge}/{\rm M}_\odot$	measured $M_{\rm BH}/{\rm M}_\odot$
A	$(4.4 \pm 1.6) \times 10^9$	$(6.6 \pm 0.9) \times 10^6$
B	$(4.5 \pm 1.7) \times 10^{10}$	$(6.0 \pm 1.4) \times 10^6$
C	$(3.6 \pm 1.5) \times 10^8$	$(1.1 \pm 0.5) \times 10^6$

There are two popular explanations for the relationships discussed in this section. One is that the galaxy-feedback processes we discussed in *Cosmology* Chapter 11 act to keep the galaxy bulge and black-hole masses closely matched, with a cycle of black-hole and stellar outflows simultaneously suppressing both star formation and the accretion that controlled further black-hole growth. An alternate hypothesis is that repeated galaxy and black-hole mergers help to average out mismatches and tighten the correlations.

The tight relationships between black-hole mass and host-galaxy properties in the local Universe give us a way to investigate how black holes may have grown. Figure 1.14 shows the local relationship between black-hole mass and stellar mass, together with estimated properties for two high-redshift samples of black holes in active galaxies.

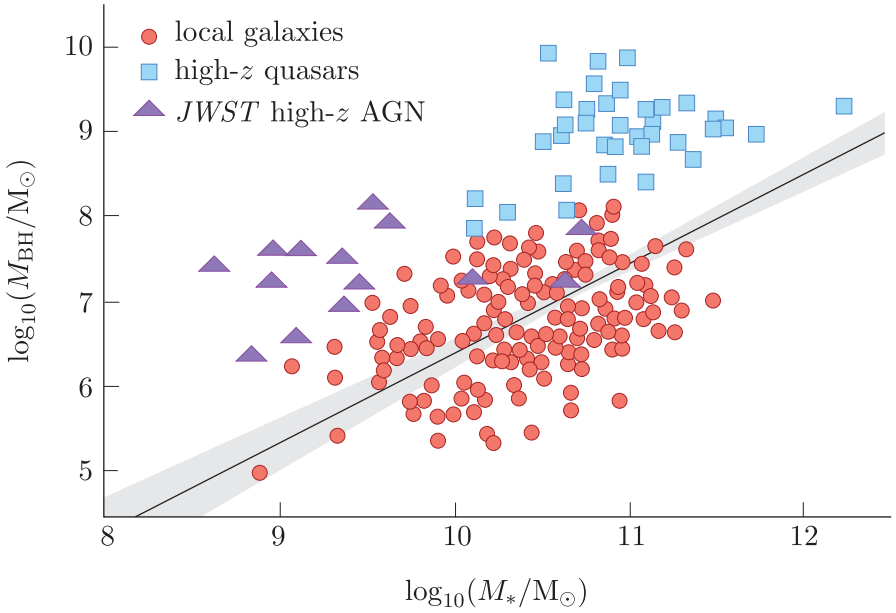


Figure 1.14 Black-hole mass and stellar mass for local galaxies (small red circles and best-fit line) and high-redshift samples (blue squares and purple triangles).

- How do the properties of the high-redshift black holes compare with the local galaxies, and what might this suggest about black-hole evolution?
- The purple triangles and blue squares are significantly above the low-redshift (local galaxy–black-hole) relationship, indicating that the high-redshift black holes are overmassive relative to the quantity of stars in their host galaxy. This may suggest that the black holes grow earlier than much of the stellar content of galaxies.

It is important to note that there are some caveats on measurements of black-hole and galaxy properties at high redshifts. The biggest challenge is what are known as **selection effects**: the objects that are easiest to observe are typically the brightest galaxies or AGN at that redshift, which means they may not be typical, and could be giving a somewhat biased view. At the time of writing, *JWST* is making many new discoveries in this area, and so we will learn more over the next few years.

1.4 Growth of black holes in the early Universe

The formation of SMBHs in galaxy centres is a long-standing puzzle in astronomy. Stellar-mass black holes (i.e. those with masses $\sim$10–150 $M_\odot$) are thought to be produced in supernova explosions at the end of the lives of massive stars. There is no single, fully understood process to produce black holes with masses $\sim 10^6$–10^9 $M_\odot$.

However, observations show that such black holes are present less than a billion years after the big bang, and that galaxies have SMBHs at their centres by the time most of their stars have formed. In this section we will explore the possible routes to grow SMBHs in the early Universe, and how they compare with observations.

1.4.1 How fast can black holes grow?

There are two possible routes for producing an SMBH: either it is 'born' supermassive or it originated as a much smaller black hole (known as a **black-hole seed**) and grew over time. The two ways that a black-hole seed can grow are by accreting matter that falls in or by merging with other black holes. Let's consider each idea in turn.

Accretion

The accreted material that crosses the event horizon of a black hole increases the mass of the black hole over time. The following example explores what accretion rates would be needed to grow SMBHs in the early Universe.

Example 1.3

Consider the growth of a black hole under the assumption that accretion is limited by the requirement not to exceed the Eddington luminosity. Assume $\eta = 0.1$.

(a) Write down an equation relating black-hole mass and accretion rate.

(b) Use this equation to derive an expression for the time t for a black hole to grow from an initial seed mass M_1 to an observed mass M_2.

(c) Calculate the times taken for two seed black holes with masses (i) 10 $M_\odot$ and (ii) 1000 $M_\odot$ to *each* grow to final masses of 10^6 $M_\odot$ (i.e. similar to the Milky Way's black hole) and 10^9 $M_\odot$.

Solution

(a) In the case of accretion limited by the Eddington luminosity, the luminosity produced by accretion (Equation 1.5) and the Eddington luminosity (Equation 1.7) are the same. Therefore:

$$\eta \dot{m} c^2 = \frac{4\pi G M_{\rm BH} m_{\rm p} c}{\sigma_{\rm T}}$$

This equation can be written as an expression for $\dot{m}$:

$$\dot{m} = k_1 M_{\rm BH} \tag{1.11}$$

where $k_1 = 4\pi G m_{\rm p}/(\eta c \sigma_{\rm T}) = 7.03 \times 10^{-16}\,{\rm s}^{-1}$ combines all the constant terms.

(b) In order to derive an expression for a timescale from Equation 1.11 we need to think about how the accreting mass and the black-hole mass are related. The accretion rate is the rate at which matter falls into the black hole. The rate at which the black hole grows, $\dot{M}_{\rm BH}$, is the rate at which mass and energy cross the event horizon. But typically the kinetic energy of the infalling matter can be neglected, therefore:

$$\dot{M}_{\rm BH} = \frac{{\rm d}M_{\rm BH}}{{\rm d}t} = \dot{m}$$

and – using Equation 1.11 – we end up with a simple differential equation involving only $M_{\rm BH}$:

$$\frac{{\rm d}M_{\rm BH}}{{\rm d}t} = k_1 M_{\rm BH} \tag{1.12}$$

We can now rearrange to find an expression for the infinitesimal time interval ${\rm d}t$:

$${\rm d}t = \frac{1}{k_1}\frac{{\rm d}M_{\rm BH}}{M_{\rm BH}}$$

and so the time taken for the mass to grow from M_1 to M_2 will be

$$t = \frac{1}{k_1}\int_{M_1}^{M_2}\frac{{\rm d}M_{\rm BH}}{M_{\rm BH}} = \frac{1}{k_1}\left[\ln M_{\rm BH}\right]_{M_1}^{M_2} \tag{1.13}$$

(c) The calculated time intervals for each of the scenarios considered, as determined using Equation 1.13, are given in Table 1.2.

Table 1.2 Calculated time intervals for SMBH growth.

Seed	$M_1/{\rm M}_\odot$	$M_2/{\rm M}_\odot$	t/y
(i)	10	10^6	5.2×10^8
	10	10^9	8.3×10^8
(ii)	1000	10^6	3.1×10^8
	1000	10^9	6.2×10^8

Example 1.3 shows that, with Eddington-limited accretion, it takes timescales in the region of a billion years to grow SMBHs comparable to the largest, billion-solar-mass examples that have been observed. This timescale is comparable to (or larger than) the age of the Universe when the earliest such black holes are observed.

The example gives us insights into what models of black-hole growth are realistic. The simplest idea of growth via accretion from small (stellar-mass) black holes is compatible with the properties of the Milky Way's black hole ($M_{\rm BH} \sim 4 \times 10^6\,{\rm M}_\odot$). However, it is far from being

realistic for the SMBHs that are observed at high redshifts. Maintaining uniformly high accretion rates for periods of hundreds of millions to billions of years during the cosmic dawn era is thought to be incompatible with typical environmental conditions at that time: a range of feedback effects (from stellar explosions to the environmental impacts of AGN/SMBH activity) are expected to limit the supply of fuel onto the black hole. Thus explaining SMBHs at high redshifts remains an unsolved problem.

Part of the solution may be that accretion at rates higher than the Eddington rate (**super-Eddington accretion**) could be possible for short periods. For example, super-Eddington accretion might occur in situations where the geometry of the inflow is not spherical, and this may mean that a small proportion of black holes were able to grow to $\sim 10^9\,\mathrm{M}_\odot$ over the first billion years of the Universe's history.

Mergers

It is plausible to imagine that if many stellar-mass black holes were present in star clusters during early star formation, then they might merge to form more massive black holes.

- How many $100\,\mathrm{M}_\odot$ black-hole seeds would need to merge to make a $10^9\,\mathrm{M}_\odot$ black hole (of the type thought to be present in $z = 6$–7 quasars)?
- This would require the merger of 10^7 black-hole seeds.

The requirement for tens of millions of mergers makes it hard to explain the presence of $\sim 10^9\,\mathrm{M}_\odot$ black holes in $z = 6$–7 quasars. This is because during the formation of Population III stars, the lack of fragmentation meant that clusters of tens of millions of stars should not form at all. During the formation of later generations of stars, where larger numbers are expected to form in a given cluster, only a small fraction of stars would be sufficiently massive to end their lives as black holes. There is a possible role, however, for rapid black-hole mergers in early protogalaxies to create black-hole seeds of intermediate masses up to $\sim 10^4\,\mathrm{M}_\odot$.

In the next section we will consider the most popular current theories for the formation and growth of black holes, as well as how it is hoped they can be tested in future.

1.4.2 Black-hole seeds – theory and observations

The simplest explanation for the origin of the black holes found in the centres of present-day galaxies is that they formed in supernova explosions at the end of the lives of Population III stars. Other possibilities include invoking **primordial black holes**, created immediately after the big bang, and black-hole seeds formed via **direct collapse**, in which a massive gas cloud (e.g. 10^4–$10^5\,\mathrm{M}_\odot$) collapsed into a black hole without first forming stars.

Figure 1.15 illustrates two competing theories for the origin of galaxy-centre SMBHs: a 'heavy-seed' model, in which the black holes formed with masses $\sim 10^4\,M_\odot$, and a 'light-seed' model where they formed with masses $\sim 100\,M_\odot$. In both cases, mergers and accretion must together grow the black holes to reach their present-day mass.

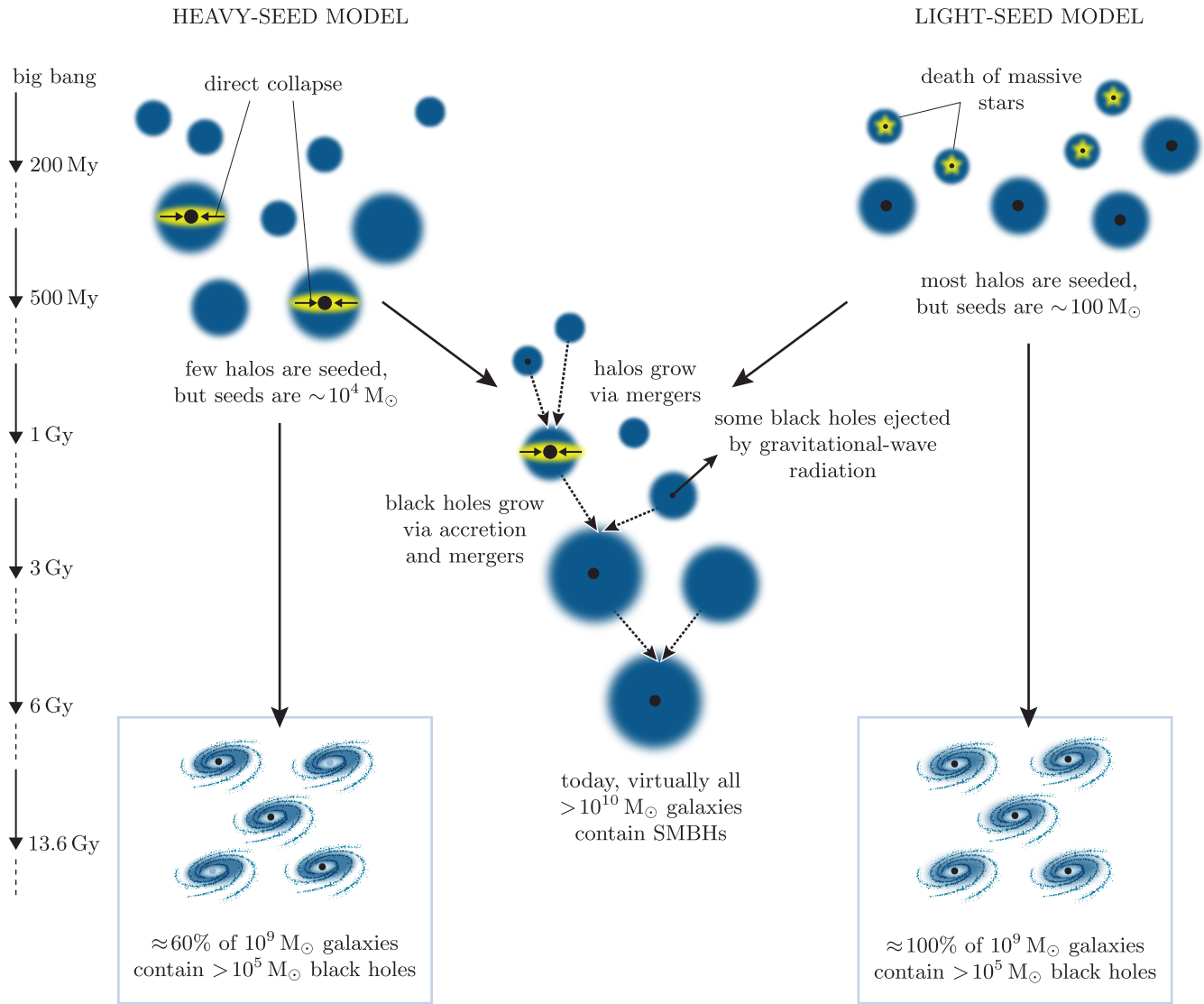


Figure 1.15 A schematic of two popular models for the origin of supermassive black holes in present-day galaxies: heavy seeds from collapse of large gas clouds (left), and light seeds formed from Population III stars (right).

This diagram suggests that the lowest-mass galaxies are a good place to look to test these theories of SMBH formation, because in heavy-seed models we might expect that some fraction of low-mass galaxies would never grow a SMBH (see the percentages given in the boxes at the bottom of each column of Figure 1.15). However, it is observationally difficult to identify black holes of $\sim 10^5\,M_\odot$.

Another route is to examine black-hole masses at the highest possible redshifts, for which *JWST* is very well suited. Figure 1.16 illustrates the possible growth histories (see the various coloured regions) for a candidate black hole that was recently identified in a galaxy at $z \approx 11$. Its black hole has been estimated via emission linewidths to have a mass of $M \sim 10^6\,\mathrm{M}_\odot$. If these results are correct, then the black hole is similar in mass to that of the Milky Way's central black hole, but exists at a time when the Universe was around 450 My old – around 3% of its current age!

At the time of writing (2023), this new result is being debated by the astronomical community. We include it here as an example of how astronomers can investigate black-hole growth with *JWST*.

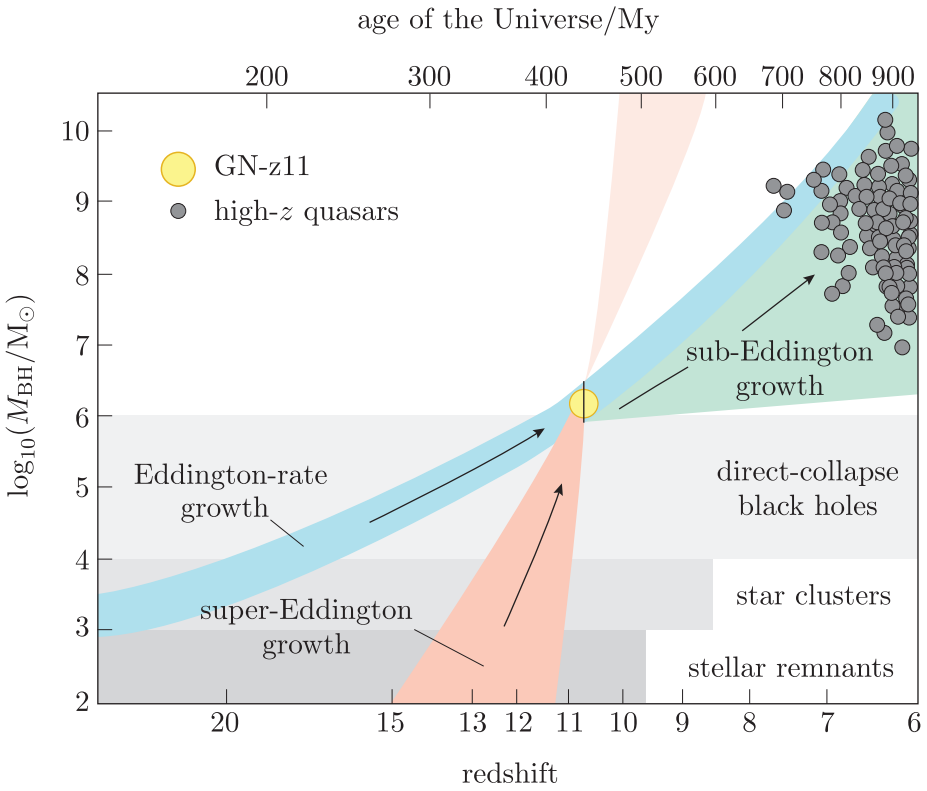


Figure 1.16 Growth histories for a candidate $10^6\,\mathrm{M}_\odot$ black hole at $z = 11$.

Use the following exercise to consider what GN-z11 may be telling us about early black-hole growth, if its inferred mass is correct.

Exercise 1.4

(a) If the earliest stars formed at $z < 20$ then, based on Figure 1.16, is it possible that the GN-z11 black hole formed from a stellar remnant?

(b) Could the black hole in GN-z11 evolve into a $10^9\,\mathrm{M}_\odot$ black hole at $z \sim 6$–7 (such as those found in some quasars shown in the figure)?

Exercise 1.4 and Figure 1.16 point to a potential role for super-Eddington accretion in helping to build the most massive black holes observed at high redshifts. The circumstances in which the Eddington rate can be exceeded are uncertain, and the subject of very active research.

An exciting future prospect to distinguish between different black-hole seed models is through gravitational wave observations. The gravitational wave detections by LIGO and the Virgo interferometer (see *Cosmology* Chapter 3) have revealed the mergers of black holes of (interestingly high) stellar masses in nearby galaxies. The future European Space Agency's *Laser Interferometer Space Antenna* (*LISA*) mission is designed to detect gravitational waves from merging *supermassive* black holes, over a large range of redshifts. *LISA* should enable us to reconstruct the histories of how massive black holes merged over time, back to the point when the first black holes became supermassive, and so reveal what the seeds for those first SMBHs were like.

1.5 Summary of Chapter 1

- The first stars are thought to have formed when the Universe was $\sim$150–200 million years old (around $z \lesssim 20$). They initiated the process of reionisation, which transformed the intergalactic medium (IGM) from a neutral to an ionised gas.
- The lack of heavy elements and the impact of this on gas cooling is likely to have resulted in larger masses for the first (Population III) stars, compared with present-day star populations.
- Soon after the first stars, as halos grew more massive and enriched with metals, the earliest galaxies were able to form and came to dominate the process of reionisation.
- Observations of the spectra of quasars reveal absorption features associated with the presence of intervening neutral gas along their light path towards Earth. The **Gunn–Peterson trough** is caused by redshifted **Lyman-α** absorption from neutral gas and is seen above $z \approx 6$, indicating that this is roughly when reionisation of the IGM was complete.
- The rate of production of ionising photons from different sources, such as massive stars and quasars, can be compared to the number of photons required to ionise a given volume of gas. This allows us to compare the roles of different ionisation sources during **cosmic dawn**.
- The earliest known galaxies have redshifts such that we are observing their properties at the time when the Universe was $\sim$400 million years old – a few per cent of its current age. They are typically first identified in deep galaxy surveys using the **Lyman-break method**, and then followed up with spectroscopy to confirm their redshift. **Gravitational lensing** can enable us to detect and study the highest-redshift galaxies in more detail.
- The population of galaxies observed at the earliest times has a steeper luminosity function, indicating relatively fewer massive galaxies (as expected from theories of galaxy evolution). This population of galaxies was typically more disc-like and irregular in structure compared with nearby galaxies.

- Supermassive black holes (SMBHs) are now being found at redshifts that are as high as those of the earliest galaxies, but the space density of quasars is much lower than galaxies at the earliest times we can observe.
- Black-hole mass is a key property to measure. It can be estimated by comparing quasar luminosity to the **Eddington limit**, the theoretical maximum luminosity for a given black-hole mass.
- Quasar luminosity is linked to the black-hole **accretion rate**, via

$$L = \eta \dot{m} c^2 \qquad \text{(Eqn 1.5)}$$

  while the Eddington luminosity is given by

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} m_{\rm p} c}{\sigma_{\rm T}} \qquad \text{(Eqn 1.7)}$$

- Black-hole mass has been found to correlate tightly with K-band luminosity, velocity dispersion, and galaxy stellar **bulge mass**, according to the relations:

$$\frac{M_{\rm BH}}{10^9\,{\rm M}_\odot} = 0.54 \left(\frac{L_{\rm K}}{10^{11} L_{\rm K\odot}} \right)^{1.2} \qquad \text{(Eqn 1.8)}$$

$$\frac{M_{\rm BH}}{10^9\,{\rm M}_\odot} = 0.31 \left(\frac{\sigma_{\rm v}}{200\,{\rm km\,s^{-1}}} \right)^{4.4} \qquad \text{(Eqn 1.9)}$$

$$\frac{M_{\rm BH}}{10^9\,{\rm M}_\odot} = 0.49 \left(\frac{M_{\rm bulge}}{10^{11}\,{\rm M}_\odot} \right)^{1.2} \qquad \text{(Eqn 1.10)}$$

  These tight correlations between black-hole and stellar properties of galaxies indicate that the growth of galaxies and their central black holes is closely linked.
- It is hard to explain the existence of the most massive black holes at high redshifts if we assume they originated from light **black-hole seeds**, which were produced by supernova collapse of the first stars and then grew via accretion at the Eddington rate (or lower).
- Alternative theories include heavy black-hole seeds produced via **direct collapse**, **primordial black holes** seeded at very early times, and/or an important role for early mergers of black holes in dense star clusters. This is an exciting topic of research with *JWST* and with the future *LISA* gravitational wave observatory.

Chapter 2 Gravitational lensing

You read in Chapter 2 of *Cosmology* that light always travels along a geodesic, which is the shortest path between any two events (or locations) in spacetime. A massive object causes spacetime in its vicinity to have a curved geometry, so light paths that travel close to such an object are bent. This phenomenon of light bending in the vicinity of massive objects can be seen in astronomical images as the effect known as **gravitational lensing**.

Figure 2.1 shows a *JWST* image of a galaxy cluster. The most striking features of this image are the elongated orange arcs that curve around the cluster's centre. Each of these arcs is a galaxy whose appearance has been distorted because of gravitational lensing. Some galaxies appear more than once, and you will find out why this happens in Section 2.1.1.

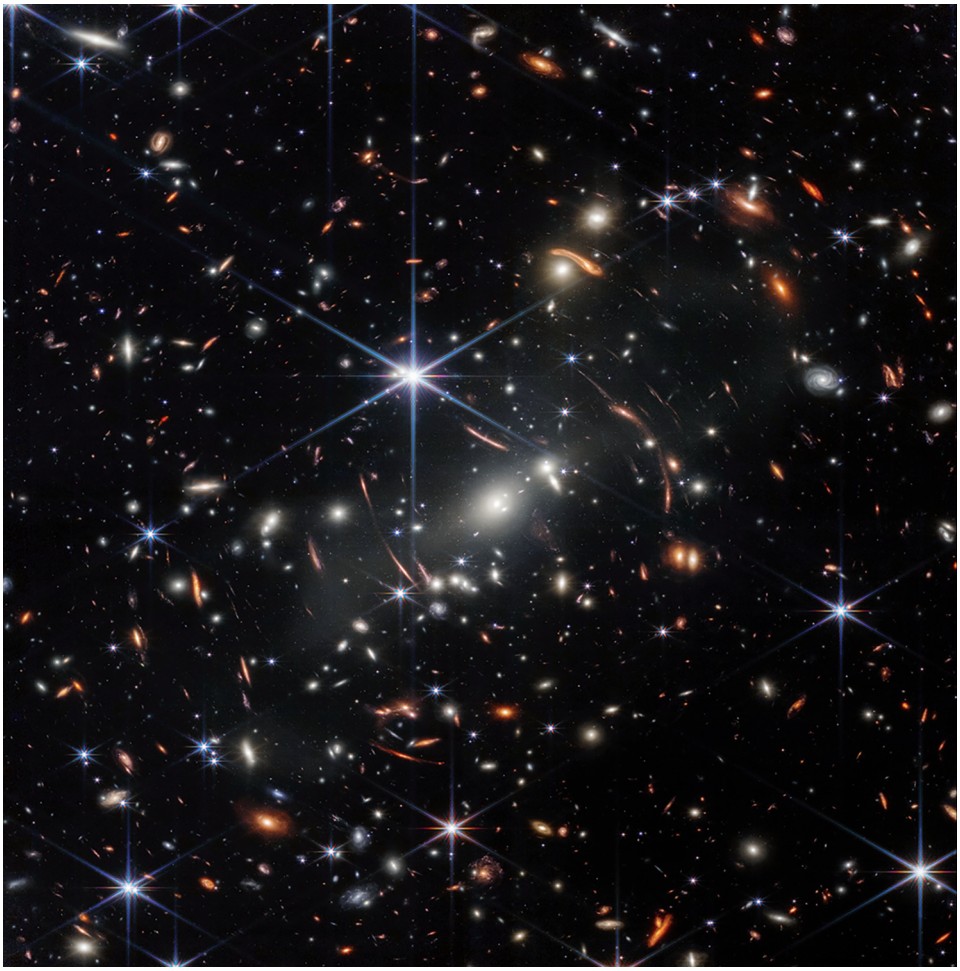


Figure 2.1 *JWST* image of gravitational lensing of the galaxy cluster SMACS 0723. This is an infrared image (a wavelength of 0.9–4.4 μm) measuring approximately 2.4 arcminutes on each side. (For reference, the Moon as viewed from the Earth is about 24 arcminutes across.)

In this chapter we will discuss how images like Figure 2.1 are formed, what they can tell us about the matter between us and a light-emitting source, and how they enable us to find and study very distant galaxies. Specifically, we will cover the three main categories of lensing:

- **strong lensing**, such as in Figure 2.1, where the distortion of images is visibly noticeable
- **weak lensing**, where image distortion is small but detectable when many galaxies are examined together
- **microlensing**, which occurs on scales too small to distinguish visually between the locations of the source and the lens, such as in lensing between stars.

Objectives

Working through this chapter will enable you to:

- understand the principles and key equations describing gravitational lensing
- carry out calculations to relate the locations and magnification of lensed images to the properties and geometry of the lensing system
- explain the difference between strong lensing, weak lensing and microlensing
- discuss the main applications of gravitational lensing for studies of the distant Universe, and key results from those studies so far.

2.1 Theory of gravitational lensing

The idea that light paths are influenced by the presence of nearby massive objects is strongly associated with general relativity. However, the notion that light paths should be bent as they pass near to massive objects dates back much earlier. Prior to the development of modern electromagnetism theory and an understanding of the nature of light, it was predicted that light particles should be accelerated by gravity according to the laws of Newton, and so have their paths deflected when passing near to a massive object like the Sun.

Einstein's theory of general relativity predicts that light travelling adjacent to an object of mass M will be deflected by an angle $\widehat{\alpha}$:

$$\widehat{\alpha} = \frac{4GM}{bc^2} \tag{2.1}$$

where b is the **impact parameter**, which is defined as the perpendicular distance of closest approach of the undeflected path to the large mass, and $\widehat{\alpha}$ has units of radians. Figure 2.2 illustrates the geometry of this situation.

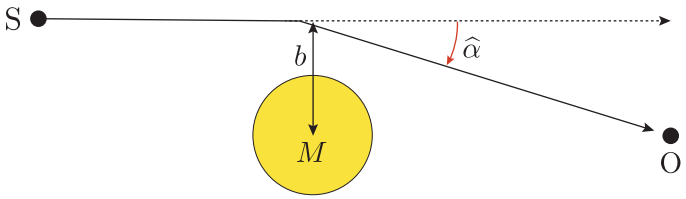


Figure 2.2 A light path travelling from a source S to an observer O is deflected by angle $\widehat{\alpha}$ as it passes a massive object of mass M.

The first measurement of this deflection was made during the 1919 solar eclipse (Figure 2.3) as part of a now-famous study by Frank Dyson, Arthur Eddington and Charles Davidson (Dyson *et al.*, 1920). This was an important early validation of general relativity.

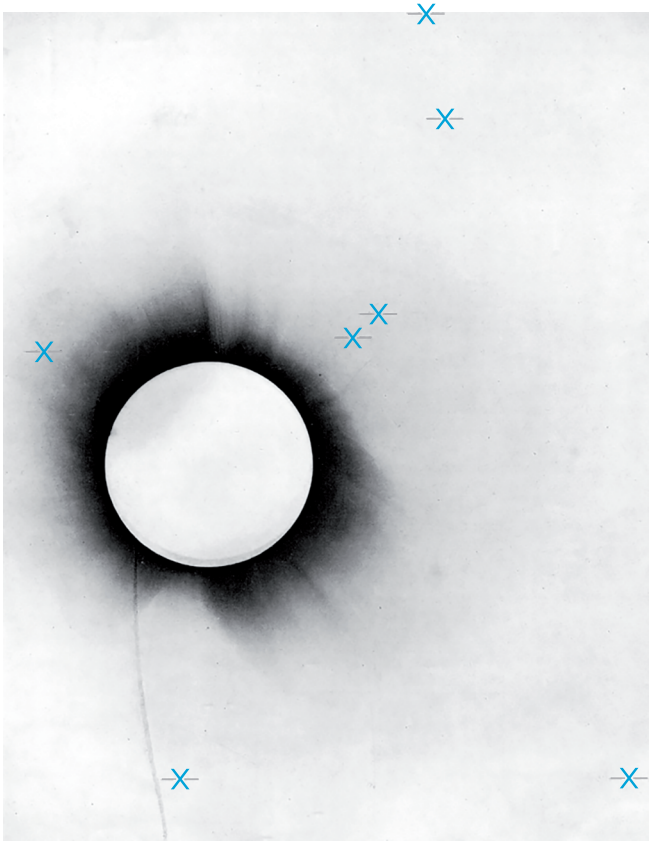


Figure 2.3 Eddington's photograph of the 1919 solar eclipse. Members of the Hyades star cluster, which were used to measure the deflection, are marked by crosses. Note that the linear arc below the Sun is an imaging artefact, not an example of gravitational lensing.

Exercise 2.1

What is the angle of deflection of a light ray travelling from a distant star, which passes at a minimum distance of $0.5\,\mathrm{R}_\odot$ from the surface of the Sun? How does this angular measure compare to the diameter of the Sun ($\sim$1800 arcseconds) or the angular resolution of Eddington's photograph ($\sim$1 arcseconds)?

2.1.1 Geometry of a lensing system

Exercise 2.1 demonstrated that a light ray passing close to the Sun exhibits a small but measurable deflection. Figure 2.4 generalises the problem, showing the light path to an observer for a source S offset from lens L (i.e. a massive object) by an angle β.

We will here consider the simplified situation in which both the lens and the source objects are point masses (i.e. all of the mass is taken to be located at a single point). In this scenario, instead of seeing the source at its true location – namely offset from the line of L and O by β – an observer at O actually perceives two deflected images of the source at apparent locations S_1 and S_2, on either side of the lens. The image on the same side of the lens as the source position, known as the primary image, (S_1 in Figure 2.4) is separated from the lens by an angle θ, and the difference between the angles θ and β is denoted α. Once again b signifies the impact parameter.

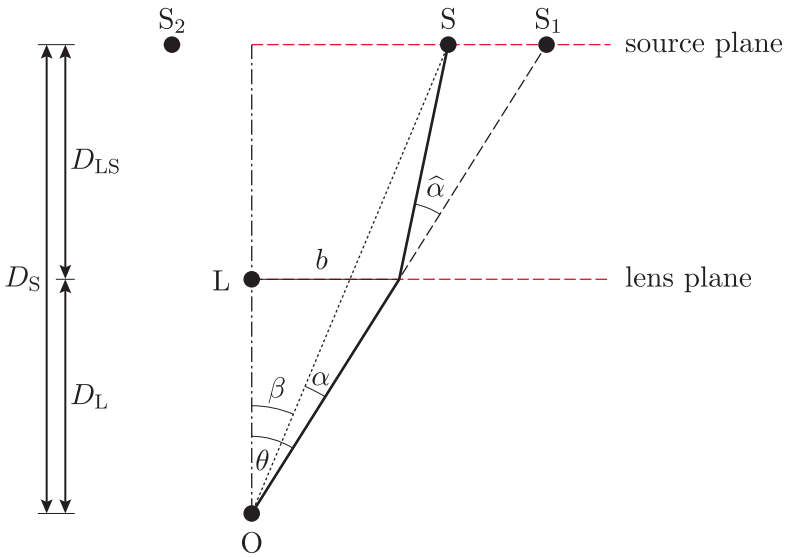


Figure 2.4 The geometry of the gravitational lensing of a source S by a lens L, with the angles that are discussed in the text labelled. The apparent positions of the source as seen from Earth are S_1 and S_2. (Light rays for S_2 are not shown, for clarity.) D_S, D_L and D_{LS} mark the distances from the observer to the source plane, the observer to the lens, and the lens to the source plane, respectively. Note that, as will be discussed in this chapter, expansion of the Universe means that for distant lenses $D_S \neq D_L + D_{LS}$.

From Figure 2.4 we can derive the **lens equation**, which relates the three key angles of the lensing geometry:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha} \tag{2.2}$$

Note that we have written the three angles, β, α and θ, as vectors. This is because each angular separation has a direction on the sky. In the situation labelled in the figure, all of the angles have the same direction of offset from the lens. However, if we apply the same definition of θ to the

secondary image (the image on the opposite side of the lens to the source, i.e. S_2) then it will have the opposite direction. By treating the angles as vectors, Equation 2.2 can be applied to both lens images. A crucial consequence of these definitions is that all angles corresponding to the side of the lens opposite to the source are assigned negative values, and so S_1 will have a positive value of θ, while S_2 will have a negative value of θ.

The thin lens approximation

A further important assumption being made here is that the physical thickness of the lens (i.e. its width in the line-of-sight direction) is very much smaller than D_L and D_{LS}, so that the bending effectively happens instantaneously. This is known as the **thin lens approximation**.

Using the lens equation

We can express the lens equation for the primary image, S_1, as a sum of the physical distances that are encompassed by each of the three angles in the **source plane** (the imagined surface perpendicular to our sight line at the distance of the source). Using Figure 2.4, and the small-angle approximation from *Cosmology* Chapter 1, we see that in the source plane:

- the lens–source offset distance is given by βD_S
- the source–S_1 distance is given by $\widehat{\alpha} D_{LS}$
- the lens–S_1 distance is given by θD_S.

Therefore, the source-plane distance between the lens and S_1 can be expressed as:

$$\theta D_S = \beta D_S + \widehat{\alpha} D_{LS} \tag{2.3}$$

- What type of distance measures are D_L, D_S and D_{LS}? (*Hint*: see *Cosmology* Chapter 5.)
- In Equation 2.3 we have related these distances to observable angles. Hence these distances must be angular diameter distances.

An important consequence of the fact that these are angular diameter distances is that, on cosmological scales, $D_L + D_{LS} \neq D_S$. The reason for this is that D_L and D_S are angular size distances as measured by an observer at the Earth, whereas D_{LS} is the angular size distance that would be measured for an observer at the location of the lens. The expansion of the Universe therefore complicates the geometry! This complication can be ignored for situations where the lens is at $z \lesssim 0.3$.

In order to apply the lens equation to real situations, we want to link the geometric quantities (angles and distances) to the lens mass. We can do this by first rearranging Equation 2.3 to obtain:

$$\widehat{\alpha} = \frac{D_S}{D_{LS}}(\theta - \beta) \tag{2.4}$$

Substituting for $\widehat{\alpha}$ using Equation 2.1 gives:

$$\frac{4GM}{bc^2} = \frac{D_S}{D_{LS}}(\theta - \beta) \tag{2.5}$$

and replacing the impact parameter b for observable quantities ($b = \theta D_L$, as shown in Figure 2.4) results in:

$$\theta - \beta = \frac{4GM}{\theta c^2}\frac{D_{LS}}{D_L D_S} \tag{2.6}$$

This equation can be written more simply as

$$\theta - \beta = \frac{\theta_E^2}{\theta} \tag{2.7}$$

where θ_E is defined as:

$$\theta_E = \sqrt{\frac{4GM}{c^2}\frac{D_{LS}}{D_L D_S}} \tag{2.8}$$

The quantity θ_E is known as the **Einstein radius**, and has a physical significance that will be explained later in this chapter.

Exercise 2.2

Use the quadratic formula to show that for a given source–lens configuration (i.e. fixed values of the three distances and α), Equation 2.7 has two solutions for θ, as indicated in Figure 2.4.

In general, β will not be measurable directly because it would require knowing the unlensed source location. However, both values of θ (the positive value corresponding to image S_1, and the negative value corresponding to image S_2) may be measurable. If the distances to the source and lens are known (typically from their redshifts) then in principle we can derive the mass of the lens, M.

The point mass assumption

It is important to note that the point mass assumption is a simplification of the true situation for lensed galaxy images. In reality the lensing mass can be distributed over a large region, and light paths may travel through the galaxy halo. The light path is bent only by the mass interior to the impact parameter, rather than by the total galaxy mass.

The following example shows how to apply the lensing equation to determine the mass of a point lens. This involves the simplifying assumption that the light paths for the primary and secondary image are bent by the same mass, which, as the highlight box above explains, is not very accurate depiction of what happens in such scenarios.

Example 2.1

A galaxy at a redshift of $z_{\rm L} = 0.010$ lenses a galaxy at a redshift of $z_{\rm S} = 0.020$. Images of the source are seen at $\theta_1 = 10.3$ arcseconds and $\theta_2 = -20.6$ arcseconds from the lens. Calculate the mass of the lens, M, in units of kg and $\mathrm{M}_\odot$. Assume that the lens can be treated as a point mass, and assume that these galaxies lie within the low-redshift limit in which the Hubble–Lemaître law (*Cosmology* Chapter 1) can be used to obtain distances.

Solution

We can write out two versions of Equation 2.6, one for θ_1 and one for θ_2:

$$\theta_1 - \beta = \frac{4GM}{\theta_1 c^2}\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}} \qquad \text{and} \qquad \theta_2 - \beta = \frac{4GM}{\theta_2 c^2}\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}}$$

Since β, the lens–source angle from the observer location, is the same for both equations, we can rearrange them both for β and equate them:

$$\theta_1 - \frac{4GM}{\theta_1 c^2}\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}} = \theta_2 - \frac{4GM}{\theta_2 c^2}\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}}$$

All of the quantities in this expression are now known, except for the lensing mass, M, and so we can rearrange to find the following expression for M:

$$M = (\theta_1 - \theta_2)\left(\frac{1}{\theta_1} - \frac{1}{\theta_2}\right)^{-1}\frac{c^2}{4G}\frac{D_{\rm L}D_{\rm S}}{D_{\rm LS}} \tag{2.9}$$

To apply this equation, we need to know the distances $D_{\rm L}$, $D_{\rm S}$, and $D_{\rm LS}$. The question tells us that we can use the Hubble–Lemaître law to obtain the distances, because the redshifts are sufficiently small that the distance estimates will be very close to the angular diameter distances.

Using $d = cz/H_0$ we find that the observer–lens distance $D_{\rm L}$ is

$$\begin{aligned}\frac{2.998 \times 10^8 \times 0.010}{67.7 \times 10^3} &= 44.3\,\mathrm{Mpc}\\ &= 1.37 \times 10^{24}\,\mathrm{m}\end{aligned}$$

the observer–source distance $D_{\rm S}$ is

$$\begin{aligned}\frac{2.998 \times 10^8 \times 0.020}{67.7 \times 10^3} &= 88.6\,\mathrm{Mpc}\\ &= 2.73 \times 10^{24}\,\mathrm{m}\end{aligned}$$

and the lens–source distance $D_{\rm LS}$ is approximately

$$\begin{aligned}D_{\rm S} - D_{\rm L} &= 44.3\,\mathrm{Mpc}\\ &= 1.37 \times 10^{24}\,\mathrm{m}\end{aligned}$$

We now have all of the quantities we need to calculate M. Converting the angles given from arseconds to radians gives $\theta_1 = 4.99 \times 10^{-5}$ rad and $\theta_2 = -9.99 \times 10^{-5}$ rad.

Substituting in all of the values into Equation 2.9, and being careful to remember the minus sign of θ_2, the mass of the galaxy lens evaluates to $M = 4.58 \times 10^{42}$ kg, which is $\approx 2.3 \times 10^{12}\,\mathrm{M}_\odot$.

Note: for higher redshift ranges, we could use `astropy.cosmology` to obtain more precise angular diameter distances, and to calculate the value of D_{LS} as an angular diameter distance measured from the lens location.

Exact source–lens alignment

The lensing geometry discussed in the previous section has an interesting special case in the situation when the source and the lens are exactly aligned. This scenario has important observational consequences.

- In the case of exact alignment, i.e. $\beta = 0$, where will the lensed images be found?
- In this situation light radiating from the source will have a symmetric distribution when it reaches the lens distance – it is not possible to define a line in the source plane between the source and the lens. This means that the light must be bent equally around the lens. Parts of the lensed image are seen at all position angles around the lens, so that there are no longer just two images in particular directions.

This type of lens image is known as an **Einstein ring**, and some examples are shown in Figure 2.5. The rings are not complete in all cases because, in reality, most sources are not point-like and may be asymmetric, so that some parts of the source are not in *exact* alignment with the lens.

The radius of the Einstein ring is the quantity θ_{E} we defined earlier (Equation 2.8). This can be shown by setting $\beta = 0$ in Equation 2.6, getting both θ terms on the same side and then taking the square root to obtain Equation 2.8.

We can also define the Einstein radius as a physical radius in the lens plane, r_{E}, if we know the distance of the lens:

$$r_{\mathrm{E}} = D_{\mathrm{L}}\theta_{\mathrm{E}} \tag{2.10}$$

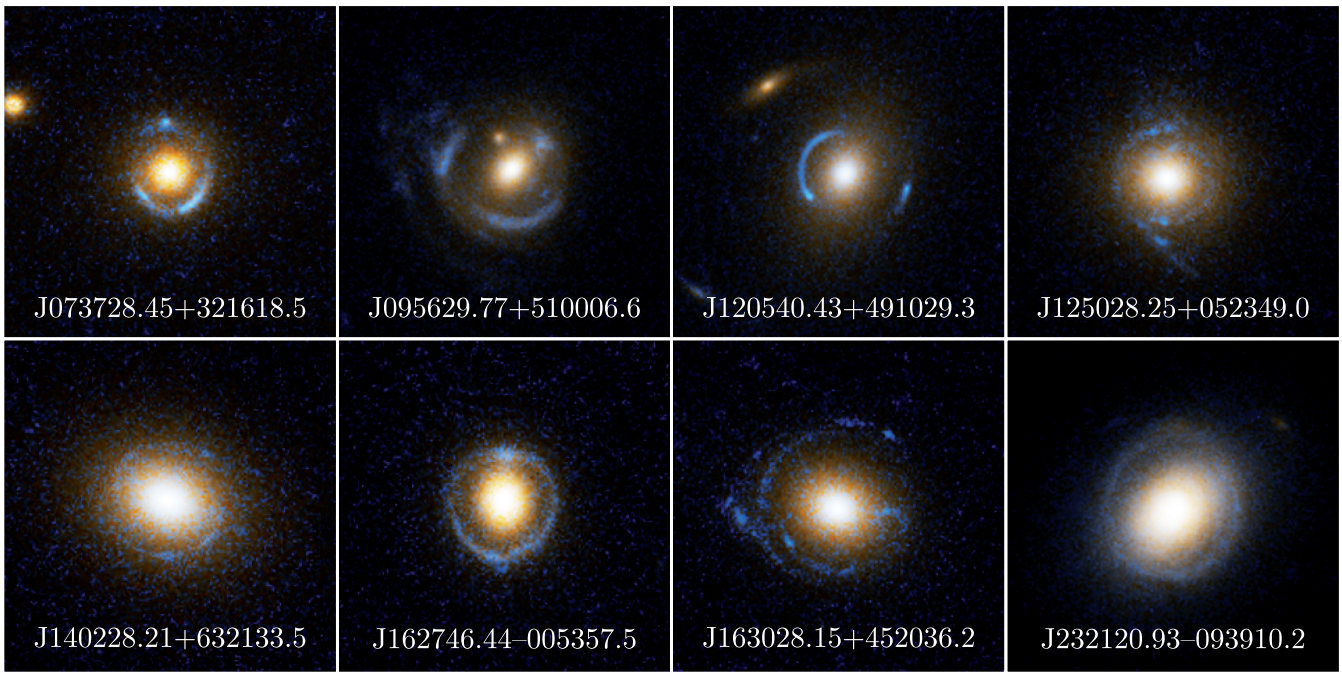


Figure 2.5 Einstein rings around galaxies observed with the *Hubble Space Telescope* (*HST*). The lens galaxies in each case are the yellow–white central elliptical shape, with the sources appearing as blueish arcs around the outside.

The next exercise allows you to explore the angular and physical size scales involved for real lensing situations.

Exercise 2.3

Calculate the Einstein radius, in units of metres, for the following:

(a) a galaxy cluster of $3.0 \times 10^{14}\,\mathrm{M}_\odot$ at 40 Mpc from Earth lenses a galaxy that is 150 Mpc from Earth

(b) a star of $1\,\mathrm{M}_\odot$ at 4 kpc from Earth lenses a star that is 8 kpc from Earth.

You may make the approximation that the lenses act as point masses for large impact parameters and, as the systems are relatively nearby, you may assume that $D_{\mathrm{LS}} \approx D_{\mathrm{S}} - D_{\mathrm{L}}$.

Most optical telescopes can resolve objects of width approximately 1 arcsecond and larger. Space-based telescopes, as well as ground-based telescopes equipped with adaptive optics, are able to image objects as small as $\sim$0.015 arcseconds. Therefore, the Einstein ring of a galaxy cluster at typical distances can be easily resolved, as in Figure 2.1. However, in the case of lenses of much lower mass – for example where there is microlensing by individual stars – the angular separation between both stars is very small, and their light is observed as a single point on the sky.

2.1.2 Magnification

As well as distorting the apparent shape and location of a distant source, gravitational lensing can also affect its apparent brightness. It is most straightforward to consider the case of an Einstein ring: in this situation light paths travelling outward from the source that would diverge in the case of no lens are instead bent towards the observer by the lens. This convergence effect means that the observer's telescope will detect a larger proportion of the light that the source emitted than would be the case if there was no lensing object. The source image is therefore magnified relative to its true brightness, in a similar effect as that produced by the optical lens in a magnifying glass.

To consider the magnification effects more quantitatively we can characterise the source–lens geometry by a parameter u, which is the expected strength of the lensing effect. It is defined as:

$$u = \beta/\theta_{\mathrm{E}} \tag{2.11}$$

When $\beta \lesssim \theta_{\mathrm{E}}$ the light will pass close to the lens, and so the lensing effect will be strong, whereas when $\beta \gg \theta_{\mathrm{E}}$ the impact of lensing will be weaker.

Strong and weak lensing

Strong lensing is defined as occurring in geometries in which β is less than the Einstein radius.

Weak lensing occurs in situations where β is significantly larger than the Einstein radius.

The **amplification** of a source's light, A, is defined as the ratio of the observed brightness relative to the unlensed expectation. In the case of a point source and point lens, the total amplification of the two lensed images (S_1 and S_2 in Figure 2.4) is given by:

$$A = \frac{u^2 + 2}{u\sqrt{u^2 + 4}} \tag{2.12}$$

- What is the amplification in the limiting cases of $u \to \infty$ and $u = 0$?
- As $u \to \infty$ (i.e. very high source–lens angular separations), $A \to 1$, and as $u \to 0$ (very small source–lens angular separations), $A \to \infty$. In other words, a source far from the lens is negligibly amplified, while a source directly behind the lens is strongly amplified.

In reality the exact limit of $A = \infty$ (i.e. an infinitely amplified source) isn't physically possible, because no source or lens is truly a point, and so perfect alignment is impossible.

Equation 2.12 can be used to calculate the amplification of a background source at different locations relative to the lens. Figure 2.6 is a **magnification map** for a point-mass lens. The colour gradient shows how the amplitude of the source behind the lens varies as a function of its location on the sky (in units of θ_{E}).

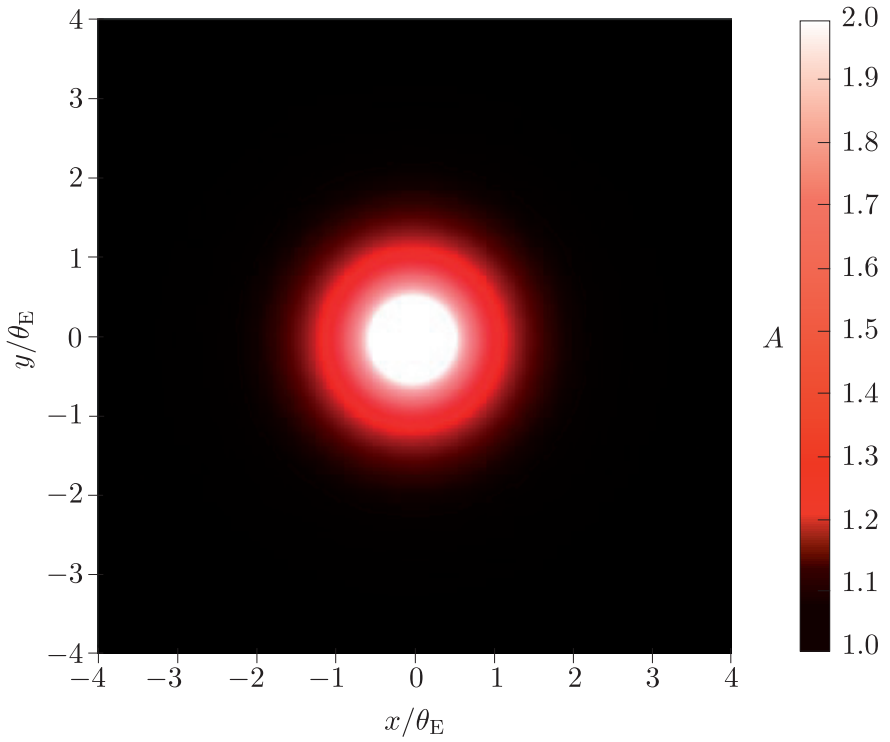


Figure 2.6 Magnification map of a point-mass gravitational lens, where all the mass is concentrated at the axes' origins.

A very important feature of magnification due to lensing is that it is **achromatic**, meaning that the amplification A is independent of wavelength. This is different from the behaviour of simple optical lenses that we use in everyday life, where the dispersion of light travelling through a medium such as glass causes wavelength-dependent effects.

In principle, the achromatic nature of all gravitational lensing enables unaltered colour information about the background source(s) to be recovered. This is most straightforward to do in the situation where both the lens and source are point-like (for example in microlensing). In situations where the source is extended on the sky, as discussed shortly, the lensing effect is still achromatic, but the wavelength-dependence of the source structure complicates the interpretation of observations.

Achromatic amplification effects can be a very useful way to confirm the *presence* of gravitational lensing. For example, microlensing signals can be distinguished via their achromaticity from other sources of variability in brightness that typically *do* depend on wavelength, e.g. stellar pulsations. The next section considers further information that can be deduced from studying microlensing.

Microlensing and the Einstein-crossing timescale

Stars move with respect to each other so, in a microlensing situation, a source star that passes behind a lensing star will move through configurations that cause different magnification for an observer. This means that the source star will seem to brighten and return to its 'normal' apparent magnitude as its light is amplified by different amounts. In these cases a characteristic light curve, like that in Figure 2.7, is produced.

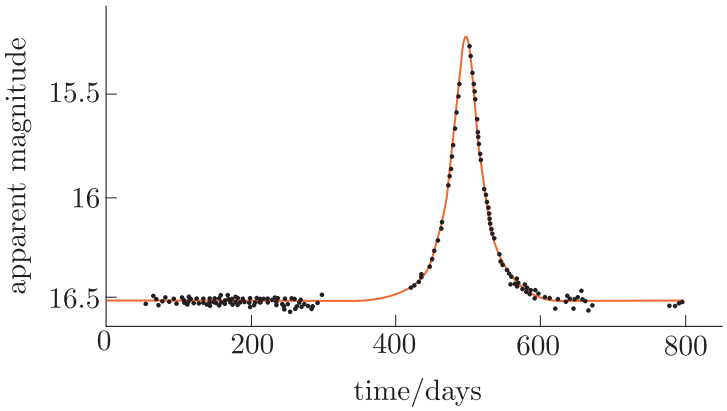


Figure 2.7 Microlensing light curve, showing the amplification as the source, lens and telescope approach perfect alignment ($u = 0$).

The amount by which the system gets brighter depends on how close the stars' approach is, and how much the source star contributes to the total flux observed for both stars.

The timescale over which the star brightens depends on the size of the Einstein ring created by the lensing (thus the mass of the lensing star and the stars' relative separations; see Equation 2.6) and on the relative velocities of the stars in the plane of the sky (their proper motions). The time taken for a star to travel a distance on the sky corresponding to the Einstein radius for the lensing geometry of the two stars is described by:

$$t_{\mathrm{E}} = r_{\mathrm{E}}/v_{\mathrm{t}} \tag{2.13}$$

where v_{t} is the transverse component of the relative velocity of the stars in the lens plane. This period is called the **Einstein-crossing timescale**.

Exercise 2.4

Assume the lens star in Exercise 2.3 part (b) has a transverse velocity of $150\,\mathrm{km\,s^{-1}}$ with respect to the source star.

(a) Calculate the timescale during which the source star passes within the Einstein radius.

(b) Comment on how the timescale varies with lens mass.

(c) What would the timescale be, in hours, if the lens were the mass of Earth ($\mathrm{M}_{\oplus} = 3 \times 10^{-6}\,\mathrm{M}_{\odot}$)?

Exercise 2.4 shows that timescales of days or even hours are needed to search for Earth-mass objects through microlensing. This is now a highly successful method to search for extrasolar planets (**exoplanets**), as will be discussed in a later section. However, similar calculations involving galaxies or galaxy clusters would show that their relative angular motions are too small to enable measurements of time variation due to changes in magnification on human timescales.

2.1.3 Extended sources and lenses

Effects of an extended source

If a source has an appreciable size compared to the Einstein radius for that lens–source system, then portions of the incident light will pass the lens with different impact parameters, b. This means that different parts of the source light will experience deflections by different angles, $\widehat{\alpha}$. The effect of this varying deflection is that the images of the source will become distorted, as well as being magnified.

Figure 2.8 shows how the images of a source are distorted as perfect alignment approaches (moving left to right across the panels), until the source eventually appears bent into an Einstein ring. The central dot represents the lens position, while the black circle is the Einstein radius. Note also the movement of the second (initially smaller and fainter) image inside the Einstein radius, and how the coming together of the two images makes the Einstein ring.

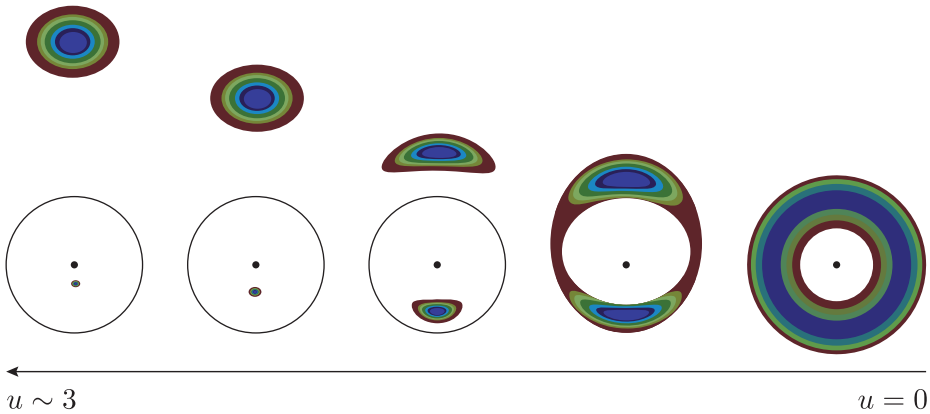


Figure 2.8 Distortion of a circularly symmetric finite source with different source–lens alignments. The value of u corresponding to the centre of the source decreases from left to right until $u = 0$ (perfect alignment), where an Einstein ring is formed (right panel). Colour is used consistently in each panel to indicate regions with the same surface brightness.

An interesting and important property of this distortion is that the surface brightness of objects is conserved: an object that is distorted over twice its nominal area will be amplified to twice its original (unlensed) flux. We can therefore use Figure 2.8 to understand an object's amplification. As each coloured region approaches the lens it is distorted over an increasingly large area; its surface brightness remains constant and so the flux is higher.

Searches for lenses therefore look for the related phenomena of image distortion and magnification. On microlensing scales, where the object's shape cannot be resolved, it is the unexpected brightening of stars that allows us to find gravitational lenses. On galaxy- and galaxy-cluster scales the simplest way to identify gravitational lenses is to look for distorted images of magnified background galaxies, as seen in Figures 2.1 and 2.5. However, it is also possible to hunt for gravitationally lensed galaxies using statistical methods to identify galaxies that appear unusually bright, implying high magnification. This can be a powerful way to identify candidate high-redshift galaxies; spectroscopic observations can subsequently be used to confirm their large distances.

Multiple lenses

As well as extended sources, another possible gravitational lensing scenario involves sources and/or lenses made up of more than one object, such as two stars in a binary system. In a situation with multiple sources we can treat each source separately, but it is worth mentioning that particularly interesting microlensing light curves can occur when the source is a binary star system. Here the two impact parameters (one for each source star) continuously change, not only due to motion of the whole system across the sky, but also due the motion of each star in the binary in its orbit.

The situation where the *lens* (rather than the source) is a binary system is more complex. The simplest case is that of a wide binary star system, where the binary orbit changes much more slowly than the Einstein-crossing time. In this situation, the microlensing light curve normally has two peaks instead of one.

The magnification map for binary lenses is also complex. For a point-mass lens, only a perfect alignment produces theoretically infinite magnification. In contrast, for a binary lens, a source will tend towards infinite magnification along a series of lines on the sky (drawn in the source plane), which form one or more **caustics**. Rather than generating an Einstein ring, light passing through these caustics produces strongly distorted images on one or more **critical curves**. We see similar effects of these caustics and critical curves from optical lenses in everyday life, for example in patterns of lighter and darker regions as light is refracted through a glass of water or onto the bottom of a swimming pool.

Figure 2.9a shows three different paths through a binary lens system comprising two unequal point masses; for example, three equally spaced source stars passing behind a lensing binary star system or, conversely, three equally spaced observers observing the same microlensing event from different observatories. (We will show a real example of a situation involving different observer viewpoints later in the chapter.) As in the single-lens case, source stars are magnified as they approach either lens. However, we can see in panel (b) that there are sharp enhancements of the magnification, A, when the paths denoted by the solid blue–grey line and the dashed green line cross over the caustic curves. Similar (though more complex) caustics exist when more than two point lenses are present.

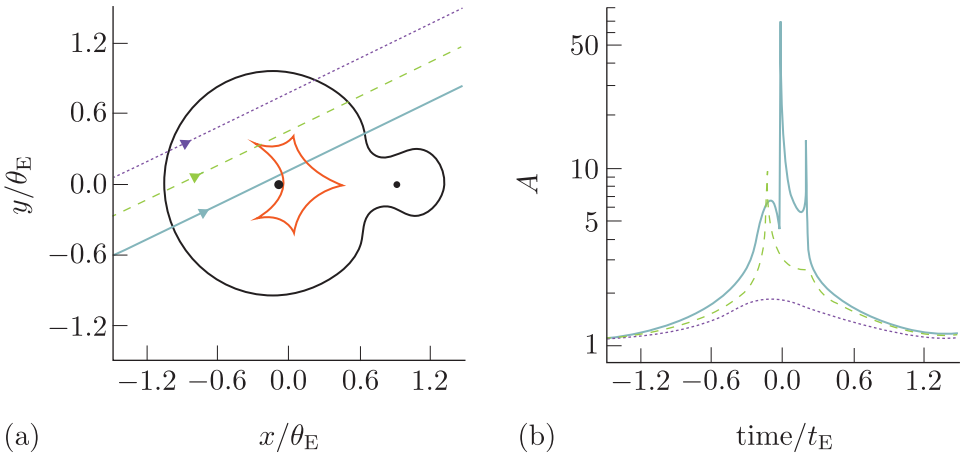


Figure 2.9 (a) Map of the positions on the sky of the critical curves (closed black curve) and caustics (closed red curve with five cusps) for an unequal binary lens consisting of two point masses (black points). (b) Light curves produced by sources traversing the diagonal lines in (a) so that the impact parameter changes with time.

We can relate the situation in Figure 2.9 to the simple point-source case by imagining that we can gradually shrink the source–lens angle to zero. As the separation tends to zero, the caustic shrinks to a point at the centre, and the critical curve about which the lensed images are arranged becomes a circle, corresponding to an Einstein ring. In this situation, the light curves in panel (b) all become single-peaked, with the solid blue–grey curve having the highest peak and the short-dashed purple curve the lowest.

Distributed-mass lenses

In cases involving galaxies, lenses can rarely be treated as points and are instead referred to as extended lenses. In this situation a light path feels the effect of only the fraction of the total lens mass contained within a spherical region of the radius corresponding to the impact parameter.

A simple example of an extended lens is an elliptical galaxy, which has an elongated distribution of mass. This arrangement results in a set of caustics for which the amplification is highest, and a set of critical curves corresponding to where the images end up.

Figure 2.10 shows a series of diagrams of the caustics and critical curves produced by such an extended lens. Panel (S) shows the source plane with two caustics centred about the lens location, and points to indicate a series of numbered locations for the source. Panel (I) shows an image of the unlensed source. The panels (1) to (8) show the lensed images that occur when the source is placed at the corresponding numbered location in panel (S). The circular and elliptical regions in the numbered panels are the critical curves around which the lensed image positions are located.

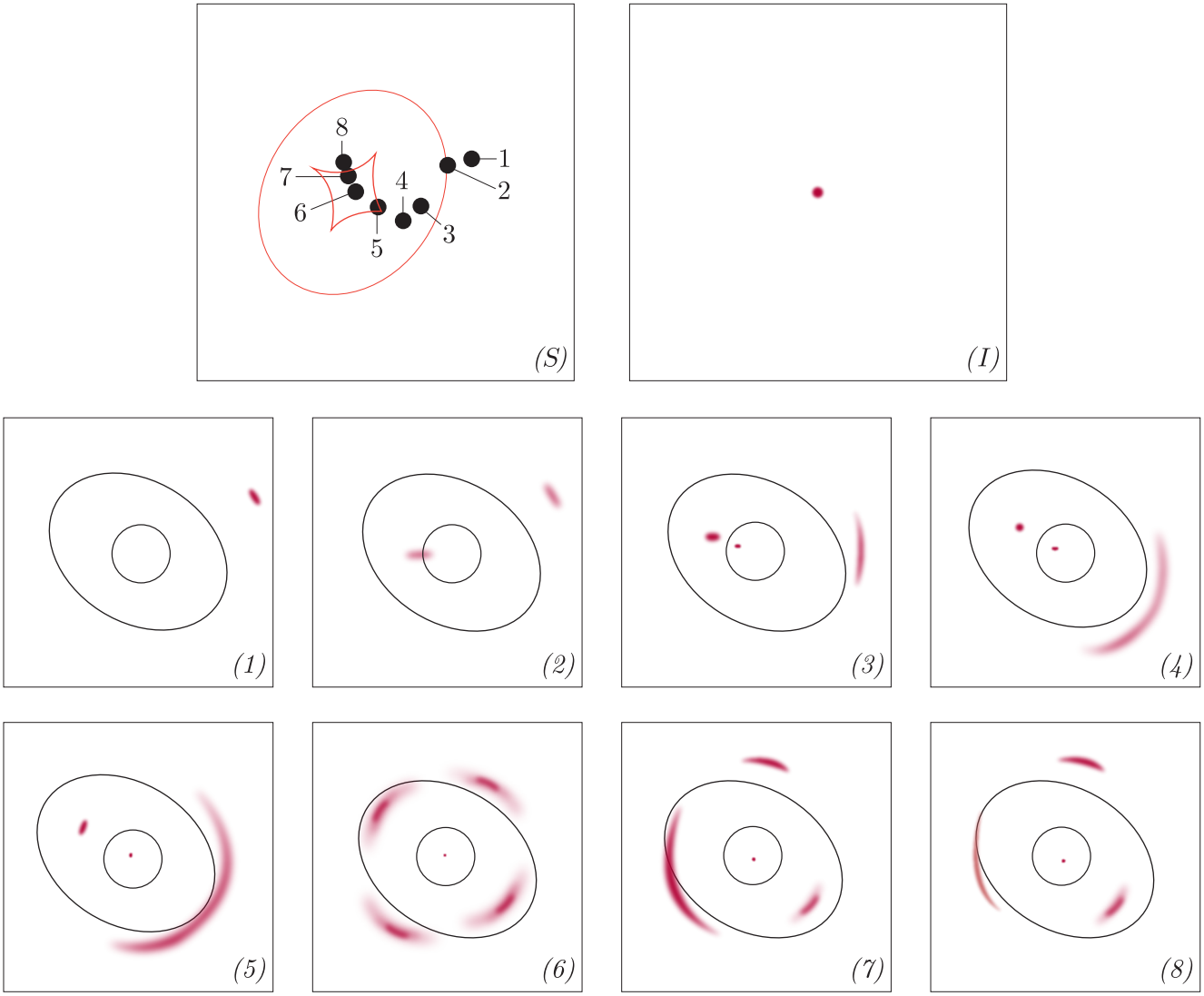
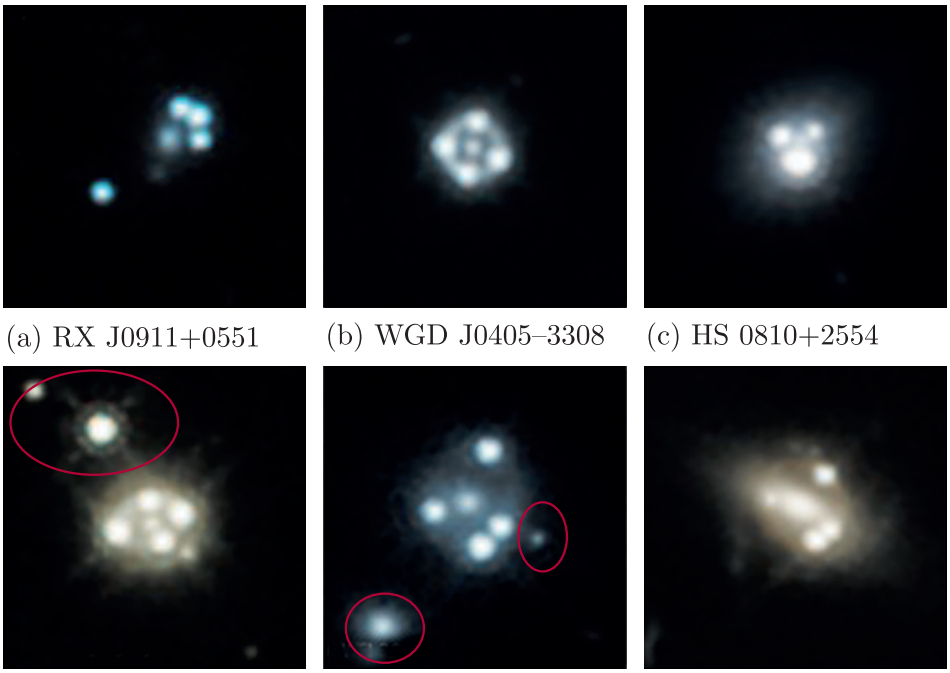


Figure 2.10 The predicted effect (panels (*1*) to (*8*)) of moving an extended circular background object (panel (*I*)) through the lens caustics caused by a simulated elliptical galaxy lens (panel (*S*)).

The various lensed image geometries shown in Figure 2.10 can be seen in real observations. Some of the most easily identifiable examples of multiple lensed image configurations caused by extended lenses involve images of distant quasars. Figure 2.11 shows some examples of quasars lensed by an extended foreground source, resulting in configurations similar to some of those in Figure 2.10 (bearing in mind that a quasar is a point source of light, unlike the extended source assumed in Figure 2.10, and so the lensed images remain point-like). Several of the panels show a configuration known as an **Einstein cross**, in which a cross-shaped configuration of four lensed images occurs.



(a) RX J0911+0551 (b) WGD J0405–3308 (c) HS 0810+2554

(d) PS J1606–2333 (e) WFI 2033–4723 (f) SDSS J1330+1810

Figure 2.11 Examples of multiple lensed images of quasars, imaged by the *HST*. Note that panels (d) and (e) include some background sources that are unrelated to the lensing – these sources are identified with red ellipses.

- Given the point-like nature of the source objects, in which panels of Figure 2.11 is the source likely to be aligned most closely with the lens?
- Panels (b) and (d) show the most symmetrical Einstein cross configurations, similar to panel (*6*) in Figure 2.10. This is the configuration in which the source is centrally aligned.

2.2 Applications of gravitational lensing

Gravitational lensing is an extremely powerful tool in astronomy and cosmology, because its measurable effects depend only on the mass and mass distribution of the lensing object or system. This makes it a unique way to search for matter that cannot be detected via the light it emits. Situations to which this applies include searching for and measuring the properties of dark matter, finding black holes, and also hunting exoplanets.

The second reason for the importance of lensing is that the magnification of light from distant objects, and the elongation of their images over larger areas, makes finding and studying very distant galaxies easier. Lensing allows us to explore galaxies at high redshift that would otherwise be too faint and/or too small in angular size to resolve details of their structure.

In the remainder of the chapter we will discuss these observational applications of gravitational lensing in more detail.

2.2.1 Microlensing

The initial application of microlensing was to investigate theories of dark matter. Historically, an important theory to explain dark matter was the existence of **massive compact halo objects** (MACHOs). These are massive solid bodies that could be orbiting in the halos of the Milky Way and other galaxies, and emit very little light. Possible MACHO populations could include brown dwarfs, neutron stars or stellar-mass black holes, and their presence can be inferred through the amplification of light from background stars that they pass in front of.

Over several decades there have been many campaigns to monitor variability in the brightness of stars in the Milky Way. The frequency at which microlensing events are detected in these surveys gives an indication of how many dark-matter objects of a given mass there are in the Galaxy. If no objects within a certain mass range are observed in a particular direction on the sky, this gives us an upper limit to the extent that objects of that mass can contribute to the total dark matter in the intervening space in that direction. For example, monitoring stars in the Magellanic Clouds to detect microlensing events allows us to place limits on the MACHO population in the Milky Way's halo, while observations in the direction of the Galactic bulge puts limits on the MACHO population in the Milky Way's disc.

Generalised limits on the fraction of dark matter that can be made of MACHOs of a given mass are shown in Figure 2.12. This plot is the result of multiple microlensing studies: the shaded regions show ranges in mass and dark-matter contribution that are ruled *out* by the observations. This evidence shows that the majority of dark matter *cannot* be made of MACHOs; microlensing is particularly important for ruling this out in objects with masses between $10^{-9}\,\mathrm{M}_\odot$ (similar to lunar mass) and $10\,\mathrm{M}_\odot$.

In the last couple of decades, evidence from the cosmic microwave background and structure formation have separately led baryonic models of dark matter to be disfavoured. Instead, the microlensing evidence against large populations of MACHOs has been used to support the currently favoured theory that dark matter is non-baryonic. However, the microlensing limits don't yet fully rule out a contribution made to dark matter from a population of primordial black holes formed in the very early Universe.

Having established that MACHOs form very little of the dark matter around us, microlensing surveys have changed their goals to finding isolated stellar-mass black holes and exoplanets. Microlensing is currently the most effective technique for finding planets with similar masses ($\mathrm{M}_\oplus$) and orbits to Earth, and free-floating planets that have become gravitationally unbound from their host stars. Figure 2.13 shows the population of known exoplanets, categorised according to their detection method.

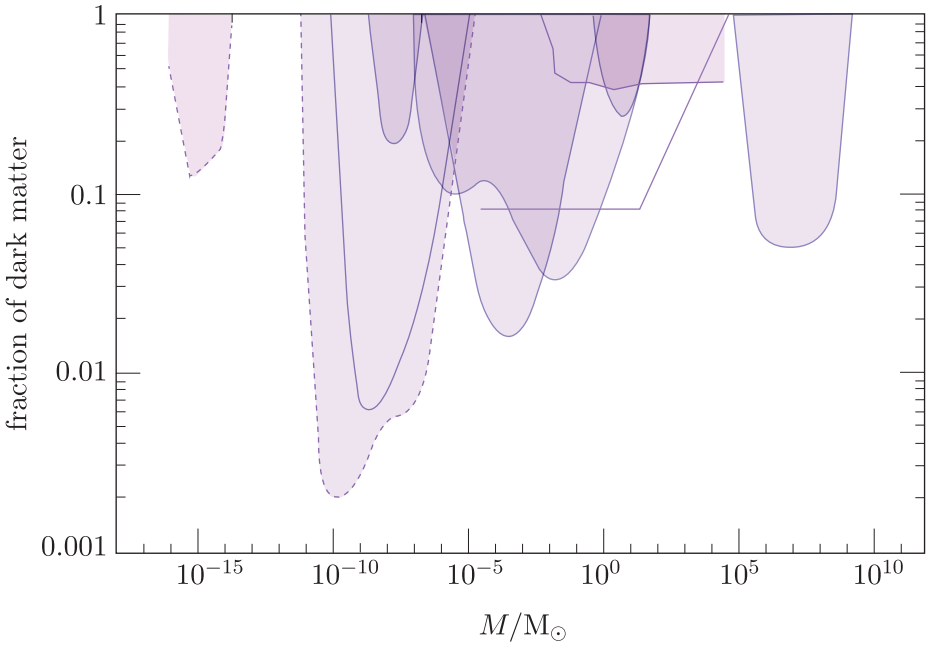


Figure 2.12 Limits from a range of microlensing experiments on the fraction of dark matter that can exist in the form of MACHOs. Dashed outlines indicate less conclusive studies; darker shaded regions indicate agreement between different studies or methods.

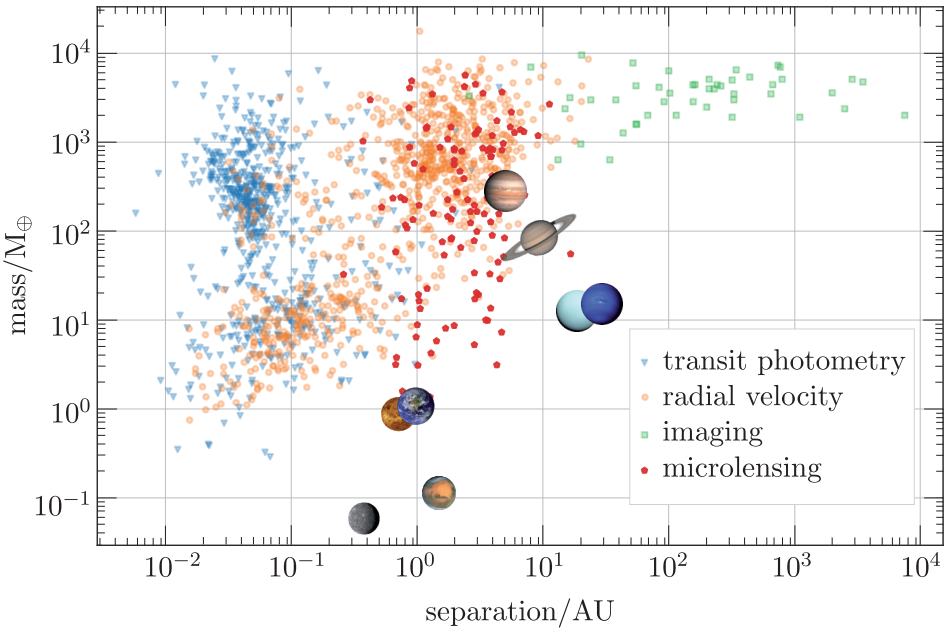


Figure 2.13 The observed exoplanet population of different masses and orbital separation (with both quantities measured relative to Earth), showing the main detection techniques for different parameter values. Microlensing is the most efficient method for detecting planets between the masses and orbital separations of Earth and Saturn, which has a mass of approximately $2\,\mathrm{M}_\oplus$ and a separation of roughly 10 AU.

Figure 2.14 shows the observed microlensing light curves of one of the microlensing detections shown in Figure 2.13: an object designated as *K2*-2016-BLG-0005Lb. Two different light curves are presented. Panel (a) shows one obtained from the *Kepler* (*K2*) *Space Telescope*, which was about 0.6 AU from Earth at the time. In contrast, the light curve in panel (b) was obtained from a collection of ground-based telescopes.

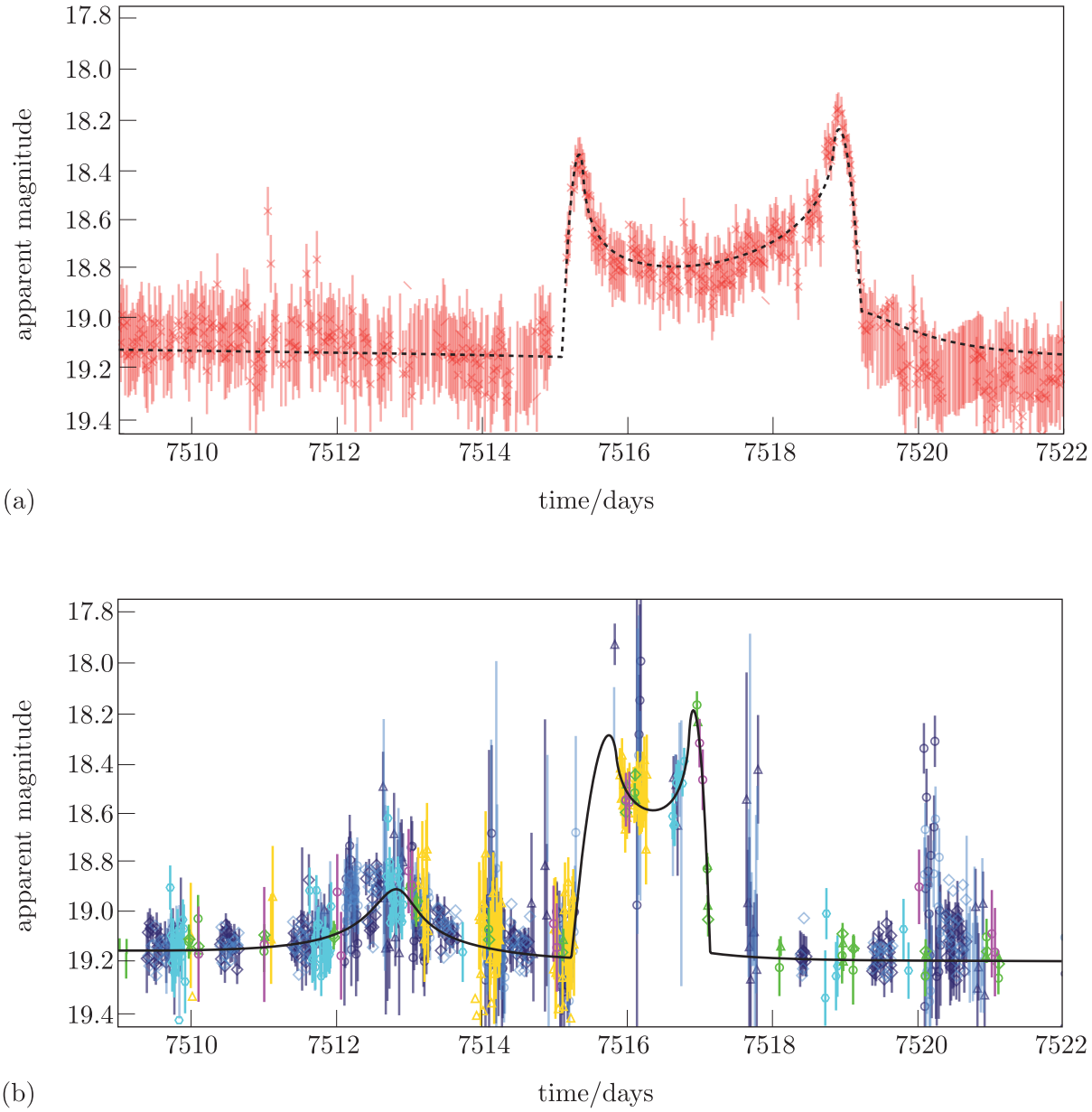


Figure 2.14 Discovery light curves of *K2*-2016-BLG-0005Lb: (a) data (and associated dotted line, which is a model fit to these data) from the *K2* spacecraft; (b) data from a number of ground-based observatories, denoted by the different colours (fitted by the solid line).

The light curves in Figure 2.14 have a number of interesting features:

- The data from Earth and space are very different from each other, which is caused by the different viewing locations of the telescopes, and lets us rule out other astrophysical explanations, like stellar flares.

- The ground-based light curve (Figure 2.14b) is achromatic. The various telescopes involved measured several wavelengths of light, but showed a consistent light-curve shape, which is a signature of microlensing.
- Both sets of data exhibit similar peaks to Figure 2.9b, indicating the crossing of a caustic feature in a binary microlens.

Subsequent modelling of the microlens relating to this exoplanet reveals this caustic to be formed by a stellar system with a star about half the Sun's mass that is orbited by a Jupiter-like planet.

2.2.2 Weak gravitational lensing

Gravitational lensing can be used as a powerful tracer of the large-scale structure of the Universe, through patterns of weak lensing. The principle works much the same as strong lensing, but instead of identifying and studying individual, clear-cut examples of magnified and distorted galaxies, it relies on the very subtle distortions that affect galaxies at larger angular distances from the lens ($u \gg 1$).

The strength of this method is that there is typically a very large number of background galaxies within a typical telescope field of observation, which means that a statistical ensemble of galaxies can be used to measure trends in the deformations of background galaxies. These deformations trace the overall mass distribution of the cosmic web. Figure 2.15 illustrates the effects of weak lensing, sometimes known as **cosmic shear**, on a large population of background galaxies. Note how the background galaxies (in blue) are subtly gravitationally lensed by the matter in front of them, which tends to make them appear aligned with the foreground cosmic web (orange).

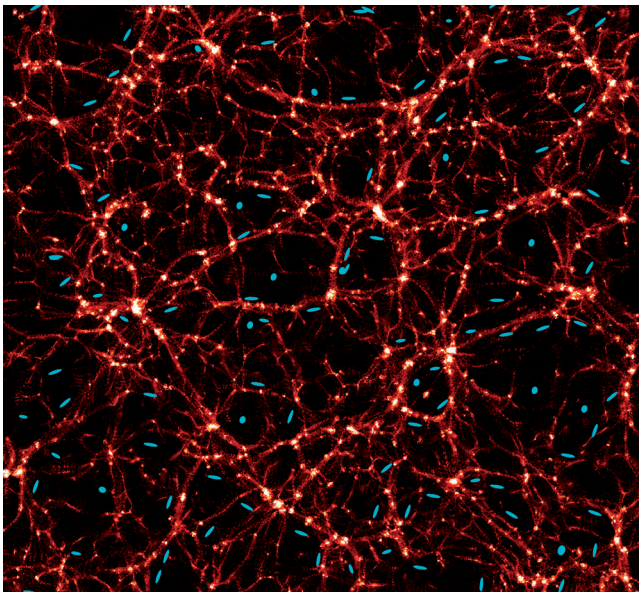


Figure 2.15 Exaggerated view of weak gravitational lensing by the cosmic large-scale structure of the Universe. Background galaxies are shown in blue, while the foreground matter distribution (the cosmic web) is shown in orange.

Weak lensing is an excellent test of cold-dark-matter (CDM) structure-formation models, and a tool for exploring how the distribution of mass in the Universe has changed over time. This also makes weak lensing useful for measuring the evolution of dark energy, and there are many ground-based and space-based imaging missions that seek to do this, with unprecedented precision. The cosmic microwave background radiation is also weakly gravitationally lensed by the intervening cosmic web, and this provides a useful consistency check of the cosmological parameter constraints from the CMB.

2.2.3 Strong lensing by galaxies

We have previously used point masses to make approximate order-of-magnitude estimates in extragalactic gravitational lensing, but we have also cautioned that the mass distributions of galaxies and galaxy clusters are *not* well approximated as point sources.

In Newtonian gravity, the force outside a spherically symmetric mass distribution is the same as if the entire mass were concentrated in a point at the centre.* According to this theory, if the Sun were suddenly squashed into a point, the Earth would continue to orbit as if nothing had happened.

Consequently, the mass that can gravitationally influence the path of a light ray is only the mass within the spherical region centred on the lensing mass whose radius is the impact parameter. For example, in a circular Einstein ring, the only mass involved is that which is contained within the ring. However, if there are several separate images of the background galaxy observed at different distances from the lens, then each image may have been gravitationally influenced by a different total enclosed mass.

A good approximation to the total mass distribution of a galaxy is often the so-called **singular isothermal sphere**, which has the same matter density profile that a self-gravitating isothermal ideal gas would have. Dark matter particles don't interact with each other so they can't be described as an ideal gas, which means it's strange that this works at all. But, if you relate the one-dimensional velocity dispersion of stars σ_v to the 'temperature' in this model, then the density profile turns out to be

$$\rho(r) = \frac{\sigma_v^2}{2\pi G}\frac{1}{r^2} \tag{2.14}$$

This model breaks down right at the centre, because the density approaches infinity as r approaches zero.

We can write the mass enclosed by a radius r as $M(r)$, which is given by

$$M(r) = \int \rho(r) 4\pi r^2 \, \mathrm{d}r \tag{2.15}$$

*This is a consequence of Gauss's theorem (sometimes called Gauss's flux theorem), which you may have met elsewhere, applied to gravity. There is a similar theorem in general relativity, called Birkhoff's theorem.

This integral evaluates to a constant multiplied by $\int \mathrm{d}r$, which means that the enclosed mass would continue to increase without limit as r increases! The singular isothermal model must therefore also break down at large radii, but can nevertheless be used to obtain useful mass estimates in appropriate radial ranges. Galaxy matter distributions also tend to have some ellipticity rather than being spherical, which further enriches the potential lensing effects.

An ellipsoidal singular isothermal mass distribution is what was used to produce Figure 2.10, which showed that the caustics and critical curves are much richer than those for point sources. If the background sources are sufficiently small in the source plane, then they can appear as four distinct images, as is seen in many lensed quasars (like those shown in Figure 2.11).

The brightness of a quasar can also vary with time, which can lead to an ingenious use of lensing to measure the local Hubble parameter, H_0. The angular diameter distances (D_L, D_S and D_LS) in Figure 2.4 are all proportional to $1/H_0$. This means that using a different value for H_0 simply scales up or scales down the whole system, as shown in Figure 2.16.

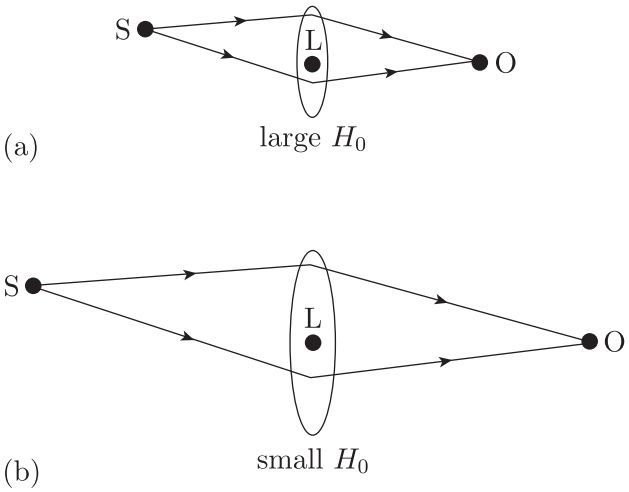


Figure 2.16 Schematic view of how the geometry of a gravitational lens changes for (a) a large, and (b) a small value of the Hubble parameter H_0. The lens is marked as L, while the source and observer are S and O, respectively. The two light paths have different lengths, and therefore different travel times.

The positions of the images in Figure 2.16 don't change with H_0, but the light travel time does change, and so will the travel time *difference* between the images. If one lensed image of a quasar flares brightly, we just need to wait for the other images to flare in order to measure the differences in light travel distance. If there is an accurate enough mass model for the lens, this will then yield a measure of H_0. This technique, known as **Shapiro delay** after its discoverer Irwin Shapiro, has also been applied to gravitationally lensed supernovae.

The properties of quasars themselves can also be explored using lensing. Lensing galaxies are made up of many individual stars, which generate complex sub-structure on top of the overall shape of the lens mass distribution, including a mesh of overlapping caustics on the scales of microarcseconds. Small objects in the source galaxy can move across these caustics, effectively providing microlenses within the larger strong lensing system. These small objects may be individual stars, or the structures around a galaxy's central black hole. The unique spectral profile of active galactic nuclei and the colour differences across their black-hole accretion discs means that the sizes of these structures can be measured via the timescale of colour changes in their lensed images.

We have skirted around *why* these unphysical singular isothermal models work at all. Unfortunately there isn't a good answer. By combining the strong and weak lensing constraints from a large sample of galaxies, it turns out that the observed stellar mass profile plus the predicted dark matter halo profile arising from CDM structure-formation theory sum up to a profile resembling a power-law, and the slope of that power-law is not too different from the isothermal model. The stellar mass dominates at small radii, while dark matter dominates at large radii. Eventually the density profile falls off more quickly than the power-law, as required by the divergence in the isothermal model.

2.2.4 Gravitational lensing with the JWST

We finish this chapter by showcasing some recent spectacular successes of using *JWST* and the most massive gravitational lenses – clusters of galaxies – to find very distant galaxies via magnification effects. Here the lensing mass distributions are more complicated, and very large magnifications are more likely.

Figure 2.17 shows a *JWST* view of the lensing cluster SMACS 0723, including the 'Sparkler' galaxy, containing what may be the most distant globular clusters of stars seen to date. Globular clusters contain some of the earliest stars to form in a given galaxy, and so it is an exciting discovery to be able to distinguish these clusters in a galaxy being observed at a time when the Universe was half of its current age. The curved red arcs in this figure are other distant background galaxies, which would be much harder to detect if they weren't magnified by the foreground cluster.

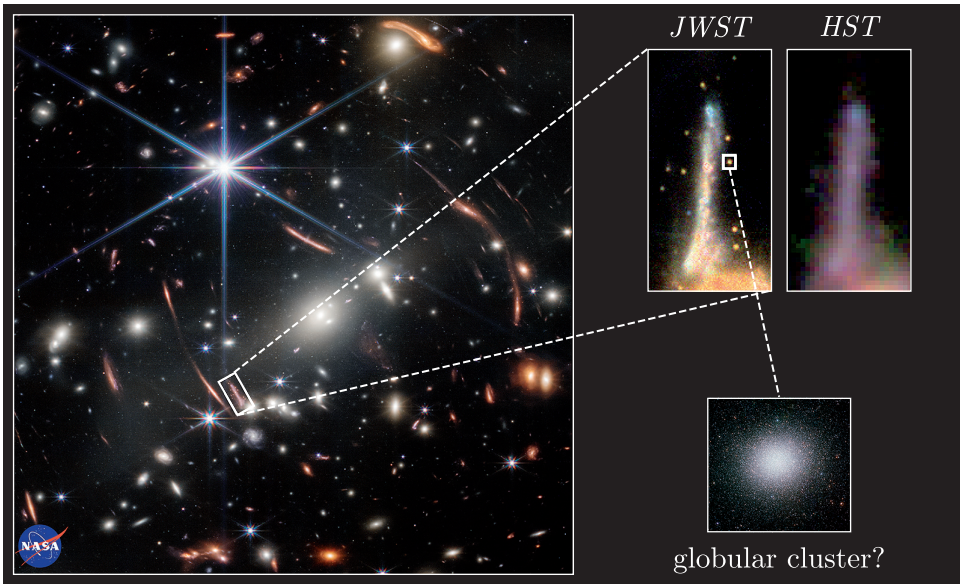


Figure 2.17 Section of a *JWST* image of the gravitationally lensing cluster SMACS 0723 (shown in full in Figure 2.1). The inset shows the Sparkler galaxy, which contains many candidate globular clusters, and compares it to the less clear *HST* image.

Figure 2.18 shows two very distant lensed galaxies discovered by *JWST*. These are candidates for some of the most distant galaxies discovered, at the time of writing.

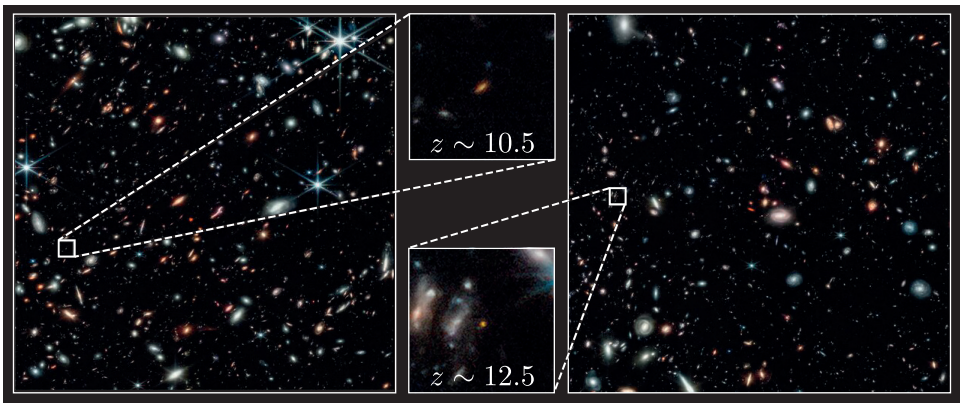


Figure 2.18 Two distant galaxies discovered by *JWST* thanks to gravitational lensing magnification by the galaxy cluster Abell 2744.

The galaxies in Figure 2.18 were predicted to be high redshift on the basis of their colours, and the most distant of these two has since been confirmed to have a redshift of $z = 12.117$. We are therefore seeing the galaxy as it was only 360 million years after the big bang. Ultimately, gravitational lensing could represent our only means to observe clusters of the earliest, Population III, stars, if such examples can be found.

2.3 Summary of Chapter 2

- A point-mass **gravitational lens** will deflect a passing light ray by

$$\widehat{\alpha} = \frac{4GM}{bc^2} \qquad \text{(Eqn 2.1)}$$

  where b is the perpendicular distance of closest approach of the undeflected light path to the lens, known as the **impact parameter**.

- Astronomical sources of light for lensing can include background stars, galaxies, quasars and even the CMB.

- Bodies that can act as lenses include planets, stars, black holes, galaxies and galaxy clusters.

- The **lens equation** for a point mass relates the angular separations on the sky between the source, the lens and the lensed images. It can be expressed as a relation between the source–image angle θ, the source–lens angle β, the lens mass M, and the angular diameter distances from the observer to the lens (D_{L}), from observer to the source (D_{S}), and from lens to source (D_{LS}):

$$\theta - \beta = \frac{4GM}{\theta c^2}\frac{D_{\mathrm{LS}}}{D_{\mathrm{L}}D_{\mathrm{S}}} \qquad \text{(Eqn 2.6)}$$

- Lensed images are magnified relative to the source brightness. For a point-mass lens, the total **amplification** is

$$A = \frac{u^2+2}{u\sqrt{u^2+4}} \qquad \text{(Eqn 2.12)}$$

  where $u = \beta/\theta_{\mathrm{E}}$ (Equation 2.11), and θ_{E} is the angular size of a lens's **Einstein ring**.

- Lensing causes the stretching and/or magnification of background source objects (or features within them), making them measurable when they would otherwise be too faint or small to be identified.

- Extended sources and lenses with extended mass distributions behave in more complex ways than point models. Extended lenses result in **caustics**, which are locations in the **source plane** where the magnification tends to infinity, and corresponding **critical curves** of highest magnification in the image plane.

- The critical curve for a point-mass lens is the Einstein ring, which has an angular radius of

$$\theta_{\mathrm{E}} = \sqrt{\frac{4GM}{c^2}\frac{D_{\mathrm{LS}}}{D_{\mathrm{L}}D_{\mathrm{S}}}} \qquad \text{(Eqn 2.8)}$$

- **Strong lensing** occurs in situations where the source has a small angular offset from the lens (i.e. $u \lesssim 1$).
  - This form of lensing is useful in measuring the properties of galaxies and galaxy clusters, especially the distribution of mass within them (including dark matter).
  - It also allows the observation of structure (and even individual stars) within high-redshift galaxies.
  - Strong lensing of time-varying sources can also provide constraints on the Hubble parameter, through the **Shapiro delay**.
- **Weak lensing** occurs in the situation of larger angular offsets between source and lens (i.e. $u \gg 1$).
  - It involves measuring the lensing distortion of a statistical ensemble of many galaxies, enabling trends in their distortion to be used as a tracer of large-scale structure.
- **Microlensing** occurs when the angular offset is too small for telescopes to distinguish between the locations of the source and lens on the sky.
  - It has been useful in determining that **massive compact halo objects** (MACHOs) – black holes or other dark, compact objects – make up *at most* a very small fraction of dark matter.
  - It is also widely used to find **exoplanets**. Unlike most other techniques, it is sensitive to relatively small planets on wide orbits as well as free-floating planets.

Chapter 3 Galaxy clusters

Galaxy groups and clusters – the gravitationally bound assemblies of tens to thousands of individual galaxies – are unique laboratories through which to study astrophysics in action. Compared to individual galaxies, clusters can be said to contain a more representative sample of the contents of the Universe: ordinary matter, dark matter, and radiation.

In *Cosmology* and in the previous chapter you saw that galaxy clusters act as nodes in the cosmic web to test how structure has evolved, and provide strong evidence for, and tools to investigate, dark matter. In this chapter we will explore galaxy clusters as astrophysical laboratories through which to study both small-scale processes of the interaction of matter and radiation, and large-scale processes of galaxy evolution.

Objectives

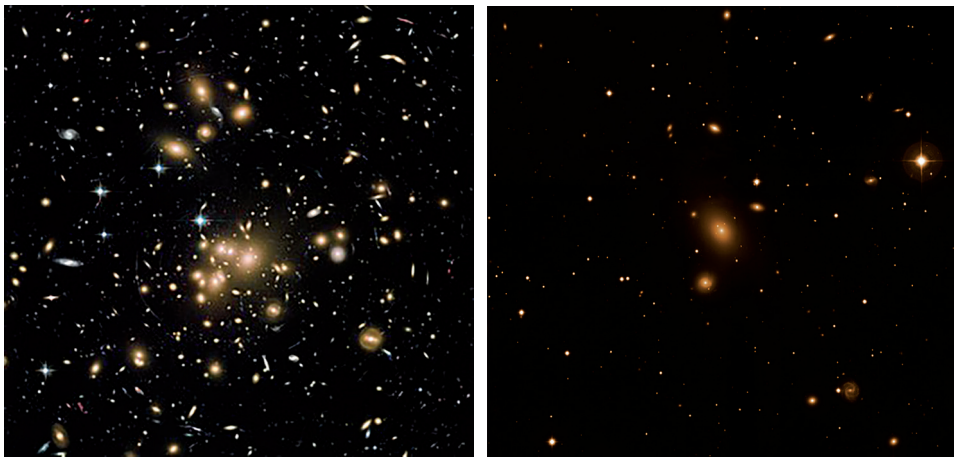
Working through this chapter will enable you to:

- summarise key methods used to identify and measure the properties of galaxy clusters
- describe the properties of the intracluster medium (ICM) and its main emission processes
- manipulate quantities and solve problems relating to X-ray emission from the ICM and the Sunyaev–Zeldovich effect
- summarise key differences between the observed properties of galaxies in clusters and those in lower density environments
- discuss the major processes influencing how galaxies in cluster environments evolve over time, including ram pressure stripping, radiative cooling and AGN feedback.

3.1 Finding and studying galaxy clusters

3.1.1 Optical and infrared observations

Galaxy clusters were identified in early telescope surveys of the sky as regions where the sky density of galaxies, i.e. the number of galaxies per square degree, was higher than the typical background distribution of galaxies. Initial investigations of individual galaxy clusters, such as the famous Coma cluster, date to the period when the nature of galaxies was first being established in the early twentieth century. The first major catalogue of around 2700 galaxy clusters was assembled by George Abell from the Palomar Sky Survey, and published in 1958. Figure 3.1 shows modern optical images of two of the clusters catalogued by Abell.



(a) (b)

Figure 3.1 Two galaxy clusters from the Abell catalogue, (a) Abell 1689 observed by the *HST*, and (b) Abell 2052 observed by the VLT. Both images show a region that spans ≈ 500 kpc.

The two clusters shown in Figure 3.1 are both sufficiently near that the observations enable the detection of all of the medium-to-large galaxies in the clusters. We can therefore conclude that the two galaxy clusters differ in their **optical richness**: Abell 1689 has a larger number and higher central density of galaxies than Abell 2052 has.

Measuring the richness of galaxy clusters is useful because it allows us to investigate how environment affects the evolution of galaxies, as well as potentially enabling cosmological tests, e.g. via the halo mass function introduced in *Cosmology* Chapter 10. Optical richness can be measured in a variety of ways. Abell counted the number of galaxies above a given luminosity threshold within a physical diameter of 2 Mpc, taking this to be a typical galaxy cluster size. Modern galaxy surveys take a variety of more sophisticated approaches including incorporating three-dimensional information, using redshift measurements to ensure that only those galaxies that are at roughly the same distance are included as cluster members.

- Would you expect the optical richness of a cluster to be related to the total cluster mass?
- The density of galaxies must be related to the total gravitational potential of the cluster: stronger gravitational forces will pull galaxies closer together. It is also logical, from the point of view of galaxy formation (see *Cosmology* Chapters 10 and 11), that more massive overdensities will form a larger number of individual galaxies. Optically rich clusters therefore have a large total mass relative to optically poor ones.

The relationship between optical richness and cluster mass is not simple, however. By the time Abell compiled his catalogue, there was already a mystery about the nature of matter in galaxy clusters. As discussed in *Cosmology* Chapter 9, Zwicky's 1933 analysis of the galaxy motions in the

Coma cluster revealed a very large **mass-to-light ratio**, so the majority of mass could not be in the form of stars. The method of gravitational lensing, discussed in Chapter 2 of this book, is another, entirely independent form of evidence for this conclusion.

Only around 3% of the mass in galaxy clusters is contained in the stars producing the optical light we observe from the constituent galaxies. Baryonic material in total makes up 10–15% of the cluster mass, but most of these baryons are not in individual galaxies: instead they are part of the **intracluster medium**, a hot X-ray emitting gas that will be discussed in the next section.

3.1.2 X-ray emission from clusters

It has been known since the 1970s that galaxy clusters are copious emitters of X-rays, which are produced across a large region spanning multiple galaxies. Figure 3.2 shows the X-ray emission measured by the *Chandra X-ray Observatory* overlaying optical images of cluster galaxies.

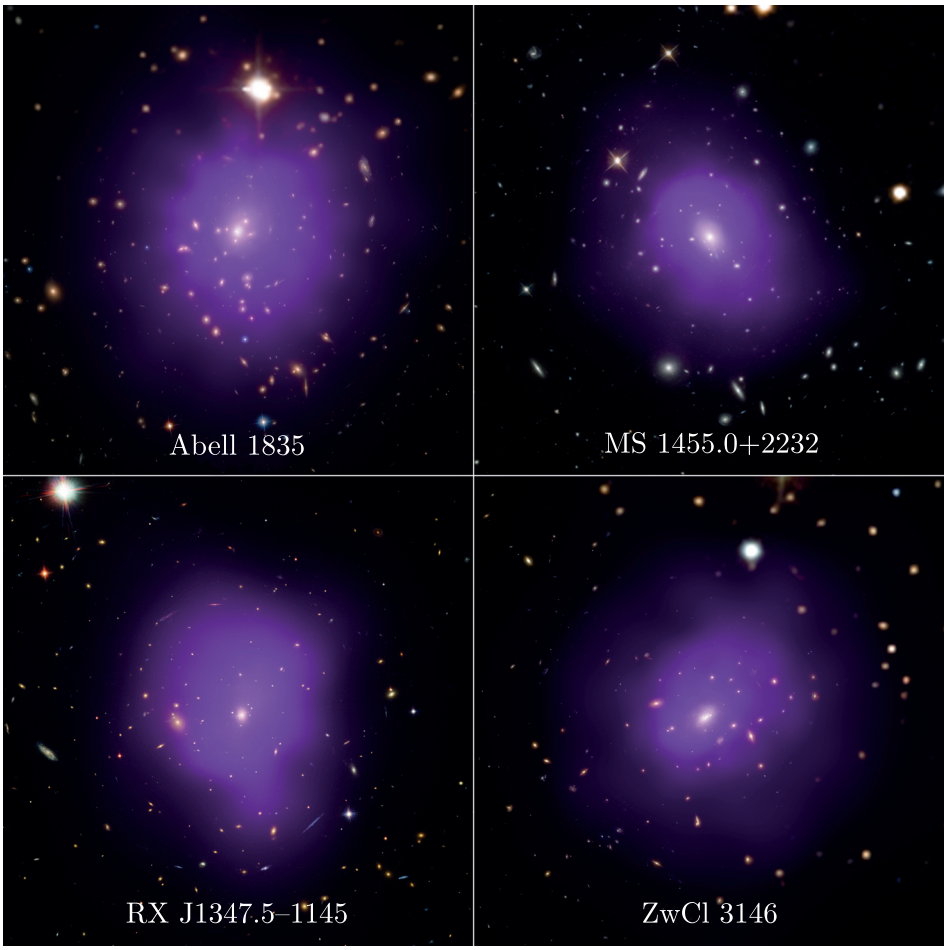


Figure 3.2 X-ray emission from four galaxy clusters (purple), as observed by *Chandra*, superimposed on their optical images.

The smooth distribution of the X-ray emission, which extends in between the individual galaxies, indicates that the radiation is not being produced by the stars and gas in individual galaxies. Instead, it reveals the presence of a low-density plasma – the intracluster medium (ICM) – pervading the entire cluster. The nature of this gas can be determined via spectroscopy: Figure 3.3 shows the X-ray spectrum of a nearby galaxy cluster.

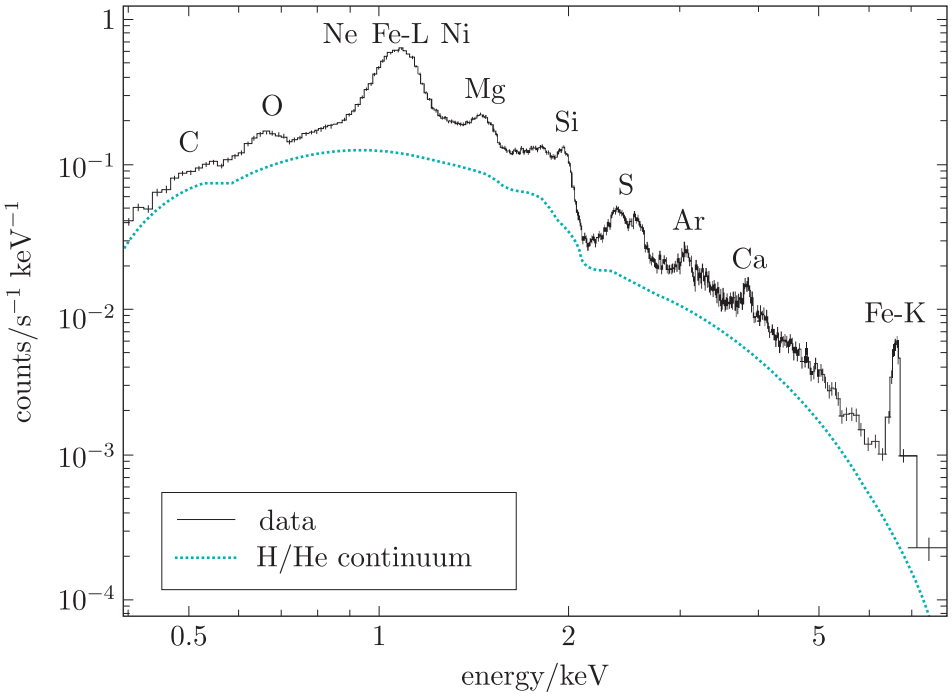


Figure 3.3 The X-ray spectrum from a region of intracluster medium in a nearby galaxy cluster.

Conventional units for X-ray and gamma-ray photon energies

You may be used to seeing the spectra of astrophysical sources expressed as functions of wavelength or frequency. In X-ray and gamma-ray astronomy, it is much more common to see spectra plotted as functions of photon energy. Conventionally, photon energies will be defined in terms of *electronvolts*.

Fortunately, the expressions for converting between a photon's energy E, its frequency ν and its wavelength λ are very simple. They involve the Planck constant, h, and the speed of light in vacuum, c:

$$E = h\nu = \frac{hc}{\lambda} \tag{3.1}$$

The shape of a galaxy cluster spectrum, such as that of Figure 3.3 is caused by two processes, **thermal bremsstrahlung**, which results in the smooth underlying continuum shape, and atomic transitions, which result in emission lines at particular energies. Thermal bremsstrahlung is a type

of radiation caused by the acceleration of charged particles as they interact with each other in an ionised plasma. Bremsstrahlung emission is also sometimes referred to as 'free–free' emission, because (in contrast to line emission processes, for example) the electrons involved remain unbound to nuclei both before and after the interaction that produces the emission.

- The X-ray spectrum is produced by ionised gas, but it is not a black-body spectrum. What does this tell us?
- Black-body radiation is produced in situations of comparatively high density: the gas must be opaque to radiation, with photons and gas particles frequently interacting. The intracluster medium is not opaque to X-ray radiation, and must have comparatively low density.

We can learn a lot about the intracluster medium from both the continuum and line emission. The **volume emissivity** of thermal bremsstrahlung emission (the power emitted per unit volume, measured in units of $\mathrm{W\,m^{-3}}$) is given by

$$\epsilon = 1.4 \times 10^{-40} g_{\mathrm{ff}} Z^2 n_{\mathrm{e}} n_{\mathrm{i}} T^{1/2} \tag{3.2}$$

where n_{e} and n_{i} are the number densities of electrons and ions, respectively, T is the gas temperature, $g_{\mathrm{ff}} \sim 1$ is the (dimensionless) Gaunt factor, a small correction factor usually close to 1, and Z is the mean ion charge (also close to 1 for a mainly hydrogen plasma).

Like black-body radiation, the shape of the continuum curve is determined by the gas temperature, which must be very high ($\sim 10^7$–$10^8\,\mathrm{K}$) to produce photons with energies that peak in the X-ray region. But Equation 3.2 shows that, unlike black-body radiation, the X-ray emission from galaxy clusters can also be used to determine gas density, which is very useful for understanding cluster physics.

The emission lines are also very interesting because they tell us about the presence of heavy elements, i.e. the astronomical metals. The hydrogen and helium in the intracluster medium are fully ionised at X-ray emitting temperatures, but heavier elements may be only partially ionised, and so can undergo electronic transitions that cause the observed lines. These lines tell us the abundances of particular elements; they are also a further diagnostic of the gas temperature and density, because the ratios between the strengths of different lines depend on those conditions.

- What can we learn by studying the abundance of astronomical metals in the ICM?
- With the exception of small quantities of lithium, all elements heavier than helium are produced by stars. The presence of significant amounts of metals in the ICM shows that it is not composed only of primordial gas produced in the big bang: chemically enriched material must have escaped galaxies in large quantities to alter the metallicity of the gas.

Detailed X-ray studies over several decades with high-resolution X-ray telescopes, including *ROSAT*, *Chandra*, *XMM-Newton* and most recently *eROSITA*, mean that the distributions of gas density, temperature and

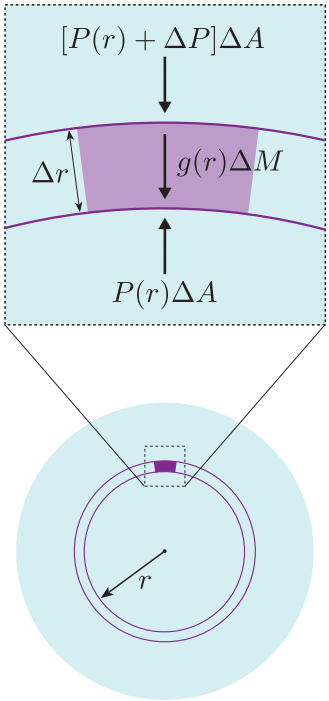


Figure 3.4 The forces acting on a shell of gas at a given radius within a gas cloud, such as a cluster atmosphere.

metallicity (including the abundances of particular metals, such as magnesium and iron) have been measured carefully for large numbers of galaxy clusters.

Hydrostatic equilibrium and cluster mass

The ability to measure gas density and temperature distributions in galaxy clusters provides a powerful way to examine their mass distributions. Just like the Earth's atmosphere, and the interiors of stars like the Sun, the intracluster medium is typically in **hydrostatic equilibrium**.

At any given location within the cluster, the pressure forces acting on a parcel of gas must balance the gravitational forces pulling it towards the centre of mass, so that the net acceleration is zero and the gas remains static. Figure 3.4 illustrates these balanced forces acting on a shell of gas at a particular radius, r, within the cluster 'atmosphere' (ICM).

The gravitational force acting on the parcel of gas with mass ΔM, located at radius r, is given by

$$F_{\text{grav}} = g(r)\,\Delta M \tag{3.3}$$

where $g(r) = GM(r)/r^2$ is the local acceleration due to gravity and $M(r)$ is the mass enclosed by a spherical shell of radius r.

The overall pressure force on the parcel of gas is the difference between the forces pushing on its inner and outer edges:

$$F_{\text{gas}} = [P(r) + \Delta P\,]\,\Delta A - P(r)\,\Delta A = \Delta P\,\Delta A \tag{3.4}$$

where $P(r)$ is the pressure at the inner edge, ΔP is the difference in gas pressure between the inner and outer edges, and ΔA is the surface area of the inner and outer edges of the purple shaded region on which the forces act (assumed to be the same, because Δr is assumed to be very small). We can express this in terms of a pressure gradient:

$$F_{\text{gas}} = \frac{\mathrm{d}P}{\mathrm{d}r}\,\Delta r\,\Delta A \tag{3.5}$$

- What happens in a situation where the pressure is the same at all radii within the atmosphere (i.e. there is no pressure gradient)?
- The net pressure force on a gas parcel will be zero, and so the gravitational force dominates and the gas falls in towards the centre.

In that situation, gravitational collapse would lead to infalling gas heating up and increasing its pressure, so that a new equilibrium would establish itself. We therefore expect a pressure gradient in the ICM with higher pressure towards the centre; this is indeed what X-ray observations show.

The equation of hydrostatic equilibrium (which also applies to planetary atmospheres and the interior of stars) can be derived by setting the sum of the two (balanced) forces to zero, and noting that $\Delta M = \rho(r)\,\Delta r\,\Delta A$, where ρ is the gas density:

$$\frac{\mathrm{d}P}{\mathrm{d}r}\frac{\Delta M}{\rho(r)} = -g\,\Delta M \tag{3.6}$$

Substituting in the previously given definition of g leads to the equation of hydrostatic equilibrium.

Hydrostatic equilibrium

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2} \tag{3.7}$$

It is important to recall that here, $M(r)$ is the *total* mass at radii less than r, and not just the gas mass. If we also remember that gas temperature and density, and therefore gas pressure (via the ideal gas law), can be measured from X-ray observations, then Equation 3.7 tells us that X-ray observations can be used to study the overall distribution of mass in clusters, just as gravitational lensing can (as discussed in Chapter 2). Example 3.1 develops this idea further.

Example 3.1

Derive an expression for how the cluster mass profile, $M(r)$, depends on the profiles of gas density, $\rho(r)$, and temperature $T(r)$, which can be measured via X-ray observations. Assume that the mean particle mass $\langle m \rangle = 0.6 m_\mathrm{p}$ (where electrons, protons and ions all contribute to this average).

Solution

First we rearrange Equation 3.7 for mass, to get

$$M(r) = -\frac{r^2}{G\rho(r)}\frac{\mathrm{d}P}{\mathrm{d}r}$$

Now we need to use the ideal gas law to relate pressure to temperature and density:

$$P = \frac{\rho(r) k_\mathrm{B} T}{\langle m \rangle}$$

An expression for the pressure gradient is obtained by differentiating using the product rule:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = \frac{k_\mathrm{B}}{\langle m \rangle}\left(\rho(r)\frac{\mathrm{d}T}{\mathrm{d}r} + T(r)\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)$$

This can now be substituted into the expression for $M(r)$:

$$M(r) = -\frac{r^2 k_\mathrm{B}}{G\rho(r)\langle m \rangle}\left(\rho(r)\frac{\mathrm{d}T}{\mathrm{d}r} + T(r)\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)$$

which can be written a little more tidily as

$$M(r) = -\frac{k_\mathrm{B} r^2}{G\langle m \rangle}\left(\frac{\mathrm{d}T}{\mathrm{d}r} + \frac{T(r)}{\rho(r)}\frac{\mathrm{d}\rho}{\mathrm{d}r}\right) \tag{3.8}$$

Equation 3.8 is commonly put into practice with modern X-ray observations; an example is shown in Figure 3.5.

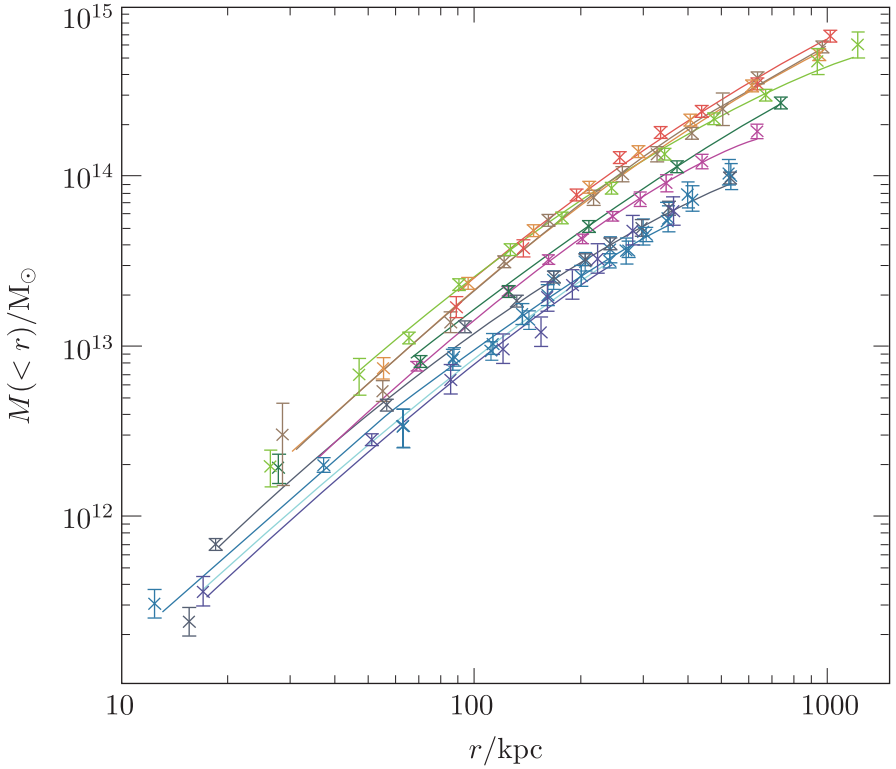


Figure 3.5 Cumulative total mass profiles for 10 galaxy clusters obtained from observations with the *XMM-Newton* X-ray observatory, using the hydrostatic equilibrium method.

Try Exercise 3.1 for some practice working with Equation 3.8.

Exercise 3.1

An X-ray observation measures a cluster to have a constant temperature of $T = 8 \times 10^7\,\mathrm{K}$ and a gas density distribution that varies with radius as

$$\rho(r) = \rho_0 \left(1 + \frac{r}{r_\mathrm{c}}\right)^{-2}$$

where $r_\mathrm{c} = 150\,\mathrm{kpc}$ and ρ_0 is an unknown constant.

Show that $M(r)$ does not depend on ρ_0 and calculate the total mass of the cluster within a radius of (a) 50 kpc and (b) 1 Mpc.

3.1.3 The Sunyaev–Zeldovich effect

Over the last decade, another method of studying the intracluster medium and galaxy-cluster mass distributions has rapidly become a very powerful tool. Millimetre-wave observations of a scattering process known as the **Sunyaev–Zeldovich effect** (named for the physicists Rashid Sunyaev and Yakov Zeldovich who first described it) provide a complementary way of finding clusters and measuring their pressure profiles.

The Sunyaev–Zeldovich (SZ) effect is a scattering process involving the photons of the cosmic microwave background (CMB), which are present at all locations and epochs across the Universe. The ionised gas particles in the intracluster medium interact with CMB photons that pass through, and energy is exchanged between them.

Exercise 3.2

Consider a CMB photon in the low redshift Universe, with $\nu = 160\,\text{GHz}$, interacting with (a) an ICM proton, and (b) an ICM electron. Compare the photon energy and particle rest-mass energies and comment on the likely direction of energy transfer.

The process involved is inverse Compton scattering (see *Cosmology* Chapter 8). The CMB photons gain a small amount of energy from the cluster ions and electrons, which means that their frequency increases (a blue shift). Figure 3.6 shows the SZ effect produced by galaxy clusters.

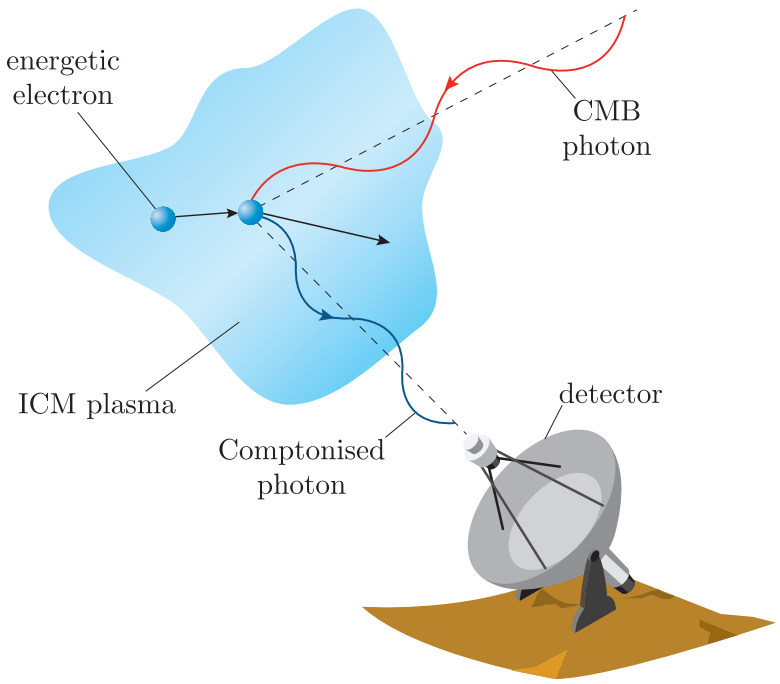


Figure 3.6 The Sunyaev–Zeldovich effect: an incoming CMB photon (red line) is scattered to higher energy (blue line) by interacting with an energetic ICM electron.

It is useful to consider the typical distance travelled by a photon passing through the centre of a cluster before undergoing a scattering interaction. This is the same as the mean free path for Thomson scattering (see *Cosmology* Chapter 1):

$$\lambda = \frac{1}{n_e \sigma_T} \tag{3.9}$$

where σ_T is the Thomson cross-section.

Typical electron number densities at cluster centres are $\sim 10^5\,\mathrm{m}^{-3}$, which leads to $\lambda \sim 5\,\mathrm{Mpc}$. This is a little larger than typical cluster diameters, but of the same order of magnitude. So the majority of CMB photons will pass through a cluster without scattering, but a significant minority – enough to produce an observationally detectable effect – will experience scattering.

The net result of the interaction for the scattered photons is a small shift in frequency, $\Delta\nu$, which is related to the typical particle energies involved in the collision:

$$\frac{\Delta\nu}{\nu} \approx \frac{k_B T}{m_e c^2} \tag{3.10}$$

Exercise 3.3

Calculate the typical frequency shift for a photon scattering off electrons in a gas with $T = 5 \times 10^7\,\mathrm{K}$. Comment on how this compares with the typical fractional deviation of cosmological anisotropies of the CMB of $\sim 10^{-5}$.

As the previous exercise shows, the SZ effect is expected to be easily detectable compared to other variations in the cosmic microwave background. Thousands of galaxy clusters have now been detected via this effect, both as a by-product of CMB missions, such as ESA *Planck*, and via dedicated experiments that can provide more detailed information.

Figure 3.7 shows maps of the SZ effect in the direction of two known galaxy clusters. The quantity that is being plotted, which is the signal measured by SZ observations, is referred to as the **Compton *y*-parameter**. It is defined as

$$y = \frac{k_B \sigma_T}{m_e c^2} \int_{l_1}^{l_2} n_e(r) T(r)\, \mathrm{d}l \tag{3.11}$$

where $n_e(r)$ and $T(r)$ are the electron number density and temperature profiles of the galaxy cluster (which may vary with radius, as discussed in Section 3.1.2) and the integral is along the line of sight through the cluster (i.e. l_1 and l_2 are the near and far edges of the cluster along our line of sight at a particular sky position).

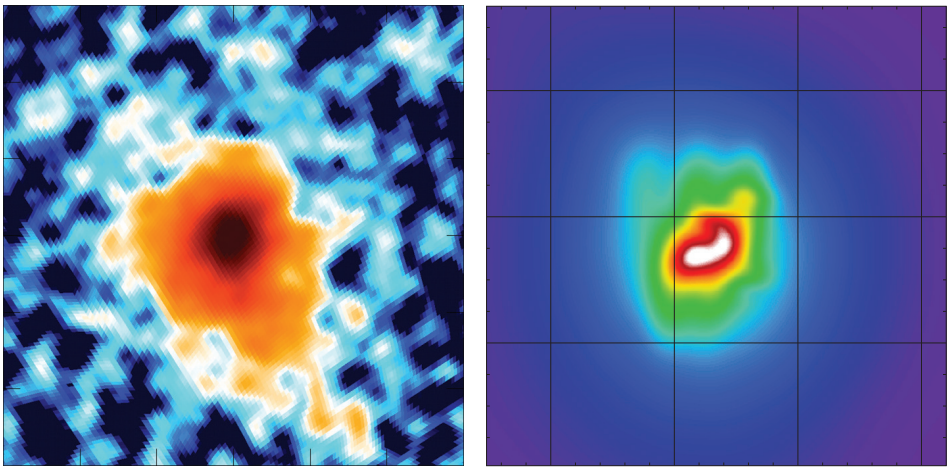


Figure 3.7 Maps of the Sunyaev–Zeldovich effect for two galaxy clusters: (a) the Coma cluster, and (b) cluster RX J1347.5–1145. In (a) blue indicates no detected SZ signal while orange to black shows an increasingly strong signal. In (b) purple corresponds to no signal, while the range of colours from turquoise to white indicates an increasingly strong SZ signal.

- What physical quantity can you infer from Equation 3.11 must be proportional to the Compton y-parameter?
- By the ideal gas law, since Compton y depends on the product of electron number density and temperature, it must be proportional to the *pressure* of the ICM gas.

One quantity that is *not* present in the formula for Compton y is the distance to the cluster, or its redshift. The strength of the dip and peak in the CMB emission caused by SZ scattering is independent of distance.

This is very different from most other signals in astronomy, where distant objects appear fainter, and is a key reason why SZ observations are important for studying galaxy clusters. SZ measurements currently provide the best way to study the most distant galaxy clusters, being able to see further into the distant Universe than X-ray telescopes.

3.2 Galaxy evolution in clusters

Observations of galaxy clusters across the electromagnetic spectrum, including X-ray and SZ observations of the intracluster medium, have greatly advanced our understanding of the environments in which galaxies evolve. In this section we will examine the differences between galaxies that evolve in dense cluster environments and isolated galaxies.

3.2.1 Comparing galaxies in different environments

Many observational studies have investigated how the properties of galaxies depend on the environment in which they are found. Figure 3.8 shows a famous relationship first identified by Dressler (1980), known as the **morphology–density relation**: the proportions of galaxies of different appearance (structural type) depend on the surrounding galaxy density, i.e. whether the galaxy is isolated or in a rich cluster.

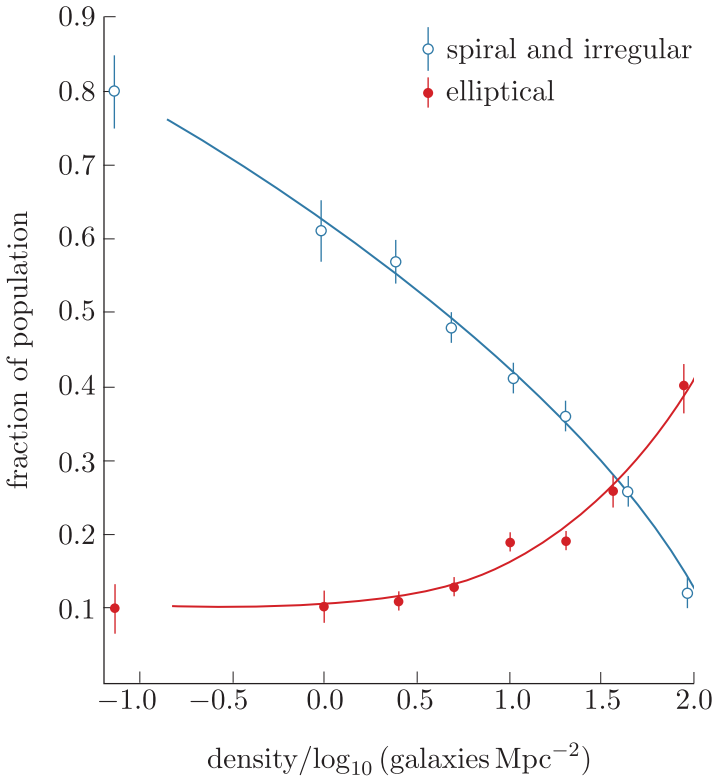


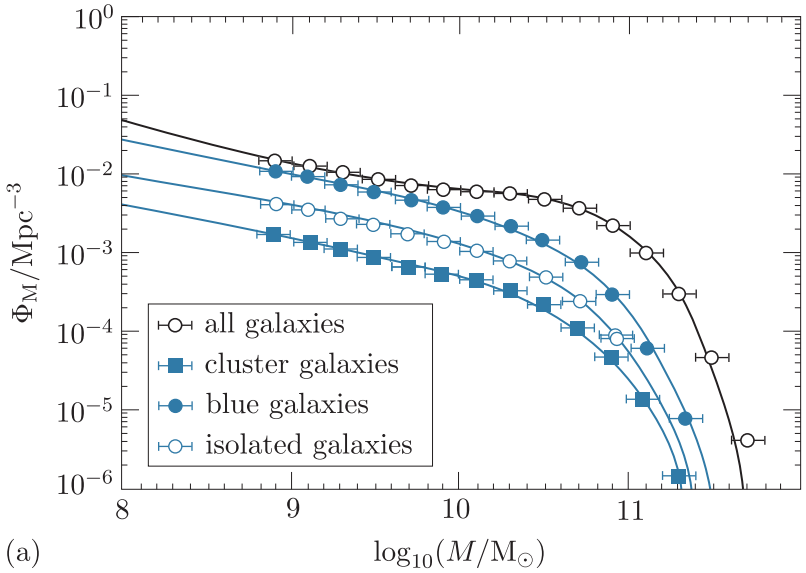
Figure 3.8 The galaxy morphology–density relation, showing how the fraction of elliptical (red) and spiral galaxies (blue) compared to the total galaxy population depends on environmental density. (Note that the plotted quantity is not a 3D density but the measured number of galaxies per unit area on the sky, which is assumed to scale with the true density.)

- What is different about the likely evolutionary histories of spiral and elliptical galaxies, and how might that depend on environment?
- A spiral structure usually forms as a result of rotation during the process of gravitational collapse to form galaxies. Elliptical galaxies are thought to be the result of galaxy mergers. The rate of such mergers is likely to be higher in environments of high galaxy density, i.e. clusters.

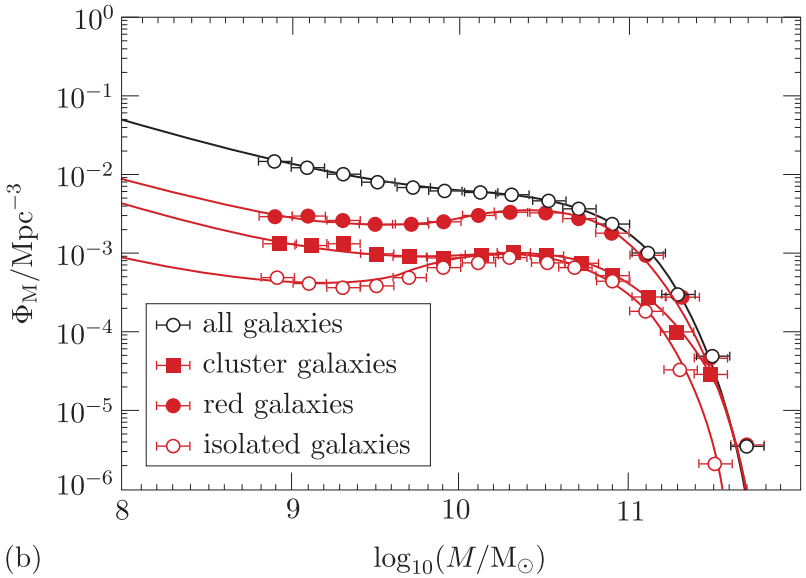
It is now well established that although both spiral and elliptical galaxies are found in clusters and in isolated environments (sometimes referred to as 'field galaxies'), the proportion of spiral galaxies decreases with environmental richness (i.e. number of near neighbours). The changes in

typical galaxy morphology go together with other differences: in particular, the rate of star formation and the quantity of gas (both molecular and atomic) that forms the fuel for star formation are both systematically lower in cluster galaxies than in isolated galaxies.

Figure 3.9 shows a comparison of the stellar mass functions for cluster and isolated galaxies, compiled from a very large survey of ∼10 000 galaxies. The solid black line in both panels is the same and indicates the total stellar mass function for all of the galaxies. The two panels show the subset of blue and red galaxies (as observed in the optical), corresponding roughly to spirals and ellipticals, with the square and circle symbols showing cluster and isolated subsets, respectively, for each panel.



(a)



(b)

Figure 3.9 The stellar mass functions for (a) blue (mainly spiral) galaxies, and (b) red (mainly elliptical) galaxies, in environments of varying richness.

Exercise 3.4

Answer the following questions based on Figure 3.9:

(a) State which type of galaxy is more common at the low-mass and high-mass ends of the mass range.

(b) Are red (mainly elliptical) galaxies in the 10^{10}–10^{11} $M_\odot$ range more commonly found in clusters or isolated environments?

(c) Are blue (mainly spiral) galaxies across all masses more commonly found in clusters or isolated environments?

(d) Is the *shape* of the mass function the same for cluster and isolated galaxies?

Research into galaxy environments, such as that shown in Figure 3.9, has shown that multiple processes must be operating to transform spiral galaxies into ellipticals. Firstly, the comparative lack of very massive spirals shows that mergers are an important part of forming the most massive galaxies. Isolated galaxies have a lower probability of merging and so are more likely to retain a spiral structure.

But mergers are not the only process that affects how galaxies evolve. The next two sections discuss (i) the effect of cluster environment and ICM on gas supply and star formation, and (ii) the effect of galaxy feedback processes linked to the central supermassive black hole.

3.2.2 Physical processes transforming galaxies in clusters

In addition to disruptive galaxy mergers caused by the higher local density of galaxies, there are several other processes associated with cluster environments that influence galaxy evolution, primarily by removing the atomic and molecular gas that provides the fuel for continued star formation. Processes that can remove gas include gravitational interaction (e.g. tidal forces), either with the cluster halo or other galaxies, hydrodynamical processes caused by interaction of the galaxy with the intracluster medium, and the suppression of infall of gas that might otherwise replenish the fuel supply for star formation.

Tidal forces as galaxies interact with each other in clusters are thought to be responsible for some stars escaping from their galaxy. One form of evidence for this process is the presence of **intracluster light** – starlight spread out through regions in between individual galaxies within clusters. Figure 3.10 shows an example of intracluster light mapped sensitively by *JWST* – the intracluster light is the smooth grey-black region encompassing the bright galaxies.

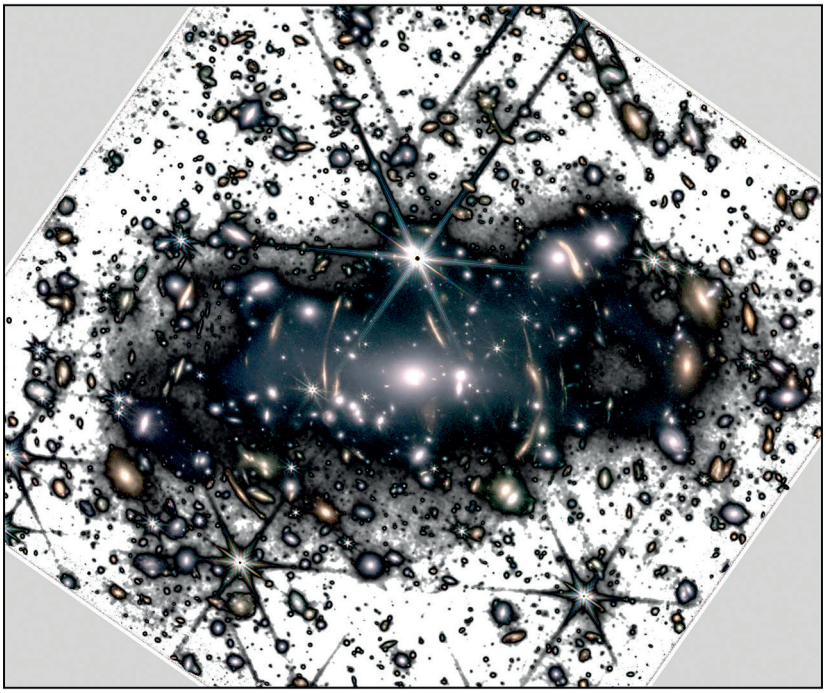


Figure 3.10 A *JWST* image of the cluster SMACS 0723 revealing intracluster light as the smooth grey-black regions between individual galaxies.

In cluster environments a process known as **ram pressure stripping** is thought to be most important for removing gas from galaxies as they fall inwards under the influence of a cluster's gravitational field and interact with its ICM. The basic idea is that as a galaxy travels through the intracluster gas it experiences a drag force, which is referred to as **ram pressure**. In some cases this force will exceed the gravitational force that binds the atomic and molecular gas to the disc of the galaxy.

The ram pressure acts in the opposite direction to the direction of galaxy motion so it can drive gas out of the galaxy, creating a characteristic tail of matter. Its magnitude, $P_{\rm ram}$, is given by

$$P_{\rm ram} = \rho v^2 \tag{3.12}$$

where ρ is the gas density at the galaxy's location within the ICM, and v is the speed at which the galaxy is travelling relative to the cluster.

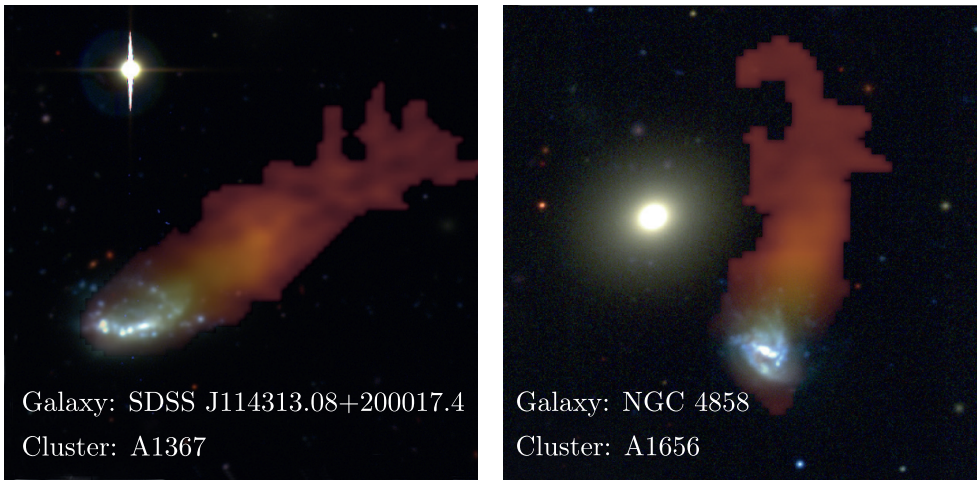
To unbind the gas from the galaxy disc, the ram pressure needs to exceed the gravitational force per unit area binding the gas to the galaxy's stellar disc, $F_{\rm grav}/A$, which can be roughly approximated as

$$\frac{F_{\rm grav}}{A} = 2\pi G \Sigma_* \Sigma_{\rm gas} \tag{3.13}$$

where Σ_* and $\Sigma_{\rm gas}$ are the **surface density** of stars and gas respectively, i.e. the mass of stars or gas per unit area. These quantities can either refer to the average surface density across the entire galaxy disc, or to localised measurements at a particular galaxy radius of interest.

- What is the average value of Σ_* (in units of $\mathrm{M}_\odot\,\mathrm{kpc}^{-2}$) for a spiral galaxy of stellar mass $M_* = 10^9\,\mathrm{M}_\odot$ and disc radius $R_* = 40\,\mathrm{kpc}$?
- Taking the mass divided by the area of (one side of) the galaxy disc (πR_*^2) gives $\Sigma_* \sim 2 \times 10^5\,\mathrm{M}_\odot\,\mathrm{kpc}^{-2}$.

Figure 3.11 illustrates ram pressure stripping in action, as measured by the International LOFAR Telescope. Two cluster galaxies are shown in the process of losing their atomic gas via this process: these objects are sometimes referred to as 'jellyfish galaxies' because of the long gas tails extending from the optical galaxy.



(a) (b)

Figure 3.11 Examples of radio-observed 'jellyfish galaxies' with extended tails of atomic hydrogen (shown in red, superimposed on optical images of the galaxies) thought to be produced via ram pressure stripping.

Example 3.2 explores the physics of ram pressure stripping in a nearby galaxy cluster.

Example 3.2

Two galaxies, NGC 4388 and NGC 4548 (which is also known as Messier 91), are both located in the Virgo cluster. Assume that the gas density profile of the Virgo cluster can be described by

$$\rho(r) = \rho_0 \left(1 + \frac{r^2}{r_\mathrm{c}^2}\right)^{-3\beta/2} \tag{3.14}$$

where r is the cluster radius, $\rho_0 = 1.7 \times 10^{-22}\,\mathrm{kg\,m^{-3}}$, $\beta = 0.5$ and $r_\mathrm{c} = 50\,\mathrm{kpc}$.

Table 3.1 lists some key properties for the two galaxies. Use this information to calculate whether it is likely that the atomic gas in each galaxy is being stripped by ram pressure at its current location.

Table 3.1 Galaxy properties. Columns are galaxy name, distance from cluster centre, estimated speed relative to the cluster, stellar mass, atomic gas mass, and radius of the stellar disc.

Galaxy	r /kpc	velocity /km s^{-1}	M_* /10^{10} M$_\odot$	$M_{\rm gas}$ /10^8 M$_\odot$	R_* /kpc
NGC 4388	363	∼1500	1.1	6.0	17.6
NGC 4548	461	∼400	4.2	4.5	12.7

Solution

To calculate the ram pressure acting on each galaxy, we first need to determine ρ at its cluster location. To do this we substitute in the given values, including the cluster radius from Table 3.1, into Equation 3.14.

For NGC 4388, at $r = 363\,\text{kpc}$

$$\rho = \left(1.7 \times 10^{-22}\,\text{kg m}^{-3}\right)\left(1 + \frac{(363\,\text{kpc})^2}{(50\,\text{kpc})^2}\right)^{-3\times0.5/2}$$
$$= 8.57 \times 10^{-24}\,\text{kg m}^{-3}$$

Note that we don't need to convert the r values to metres, since the unit conversions would cancel out. Repeating this calculation for NGC 4548, at $r = 461\,\text{kpc}$, gives $\rho = 6.02 \times 10^{-24}\,\text{kg m}^{-3}$.

We can now use the given velocity estimates to determine the ram pressure of the two galaxies using Equation 3.12.

For NGC 4388

$$P_{\rm ram} = \left(8.6 \times 10^{-24}\,\text{kg m}^{-3}\right)\left(1500 \times 10^3\,\text{m s}^{-1}\right)^2 = 1.93 \times 10^{-11}\,\text{Pa}$$

For NGC 4548

$$P_{\rm ram} = \left(6.0 \times 10^{-24}\,\text{kg m}^{-3}\right)\left(400 \times 10^3\,\text{m s}^{-1}\right)^2 = 9.63 \times 10^{-13}\,\text{Pa}$$

So the ram pressure is much higher for NGC 4388 than for NGC 4548.

Next we need to work out Σ_* and $\Sigma_{\rm gas}$ to be able to calculate the force binding the gas to the galaxy disc using Equation 3.13. We can use the same method as before, but this time converting to SI units.

For NGC 4388

$$\Sigma_* = \frac{1.1 \times 10^{10}\,\text{M}_\odot \times 1.99 \times 10^{30}\,\text{kg M}_\odot^{-1}}{\pi\left(17.6\,\text{kpc} \times 3.086 \times 10^{19}\,\text{m kpc}^{-1}\right)^2}$$
$$= 0.0236\,\text{kg m}^{-2}$$

For NGC 4548

$$\Sigma_* = \frac{4.2 \times 10^{10}\,\text{M}_\odot \times 1.99 \times 10^{30}\,\text{kg M}_\odot^{-1}}{\pi\left(12.7\,\text{kpc} \times 3.086 \times 10^{19}\,\text{m kpc}^{-1}\right)^2}$$
$$= 0.173\,\text{kg m}^{-2}$$

Repeating these calculations using the gas mass values gives $\Sigma_{\rm gas} = 1.3 \times 10^{-3}\,\text{kg m}^{-2}$ and $1.9 \times 10^{-3}\,\text{kg m}^{-2}$.

We can now calculate the binding force per unit area via Equation 3.13. These are given in Table 3.2.

Table 3.2 Comparing ram pressure and binding force for Virgo galaxies.

Galaxy	Σ_* /kg m^{-2}	$\Sigma_{\rm gas}$ /kg m^{-2}	$F_{\rm grav}/A$ /$\times 10^{-13}$ Pa	$P_{\rm ram}$ /$\times 10^{-13}$ Pa
NGC 4388	0.024	0.0013	0.13	190
NGC 4548	0.17	0.0019	1.4	9.6

For NGC 4388 the ram pressure is estimated to be more than 1000 times the binding force, and we would expect strong ram pressure stripping to operate. For NGC 4548 the ram pressure is lower and the binding force higher than for NGC 4388, but the ram pressure still outweighs the binding force, and so can remove gas from the galaxy.

3.2.3 Radiative cooling of cluster gas

Another way in which the ICM influences galaxies is via the effects of gas cooling on the central galaxy of the cluster. Many clusters have a dominant massive galaxy at the centre of the cluster, which is also typically the most optically bright, and is known as the **brightest cluster galaxy** (BCG). The cluster centre is also the location at which the X-ray emission from the ICM is strongest.

The cluster X-ray luminosity corresponds to the rate at which energy is being carried away by radiation, so it is equivalent to the rate by which the ICM gas is *losing* energy. In other words, the intracluster gas is cooling, and the rate at which energy is being lost is highest at the cluster centre.

If the energy lost to X-ray radiation is not replenished, the temperature of the central gas will decrease over time. This has important consequences for the central galaxy because a decrease in temperature corresponds to a decrease in the ICM gas pressure over time, via the ideal gas law.

- Why is the distribution of gas pressure in a cluster important?
- Hydrostatic equilibrium requires the pressure gradient of the ICM to balance the gravitational forces acting on the cluster gas at a particular radius. If pressure decreases, then gas will move inwards under the influence of gravity.

The discovery of bright, centrally peaked X-ray distributions in clusters led to the realisation that this implies a slow inward flow of rapidly cooling gas into BCGs, a phenomenon known as a **cooling flow**. Cooling flows should lead to the presence of cold molecular and atomic gas in central galaxies, as well as increased star formation.

Whether or not a cooling flow is present in a particular cluster will depend on the cooling time of the central gas, which (from *Cosmology* Chapter 11, assuming three degrees of freedom for an ionised gas) is defined as

$$t_{\rm cool} = \frac{3nk_{\rm B}T}{2\Lambda_{\rm cool}} \tag{3.15}$$

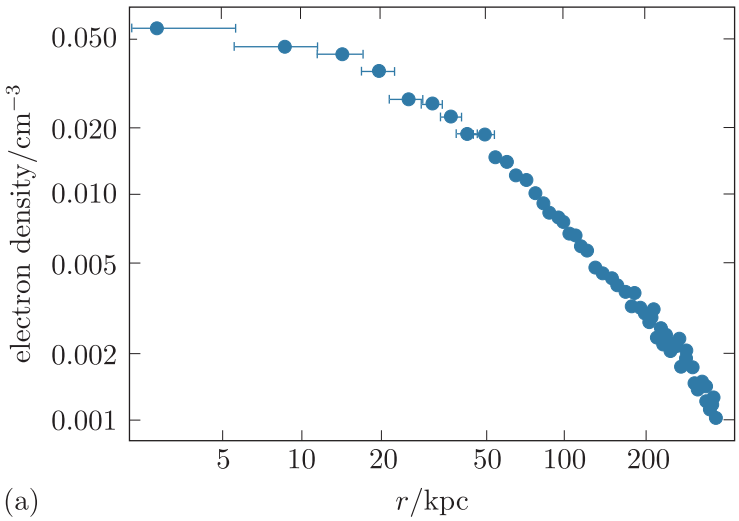
where n is the particle number density (accounting for both electrons and ions), and $\Lambda_{\rm cool}$ is the rate of energy loss per unit volume.

- If X-ray radiation is the dominant cooling pathway, how would you expect $\Lambda_{\rm cool}$ to be related to the X-ray luminosity, $L_{\rm X}$?
- The X-ray luminosity is the total energy loss per unit time, with units of $\mathrm{J\,s^{-1}}$. $\Lambda_{\rm cool}$ is the energy loss rate per unit volume, and so the cooling rate for a particular region of X-ray emitting gas is $\Lambda_{\rm cool} = L_{\rm X}/V$, where V is the volume of the region considered.

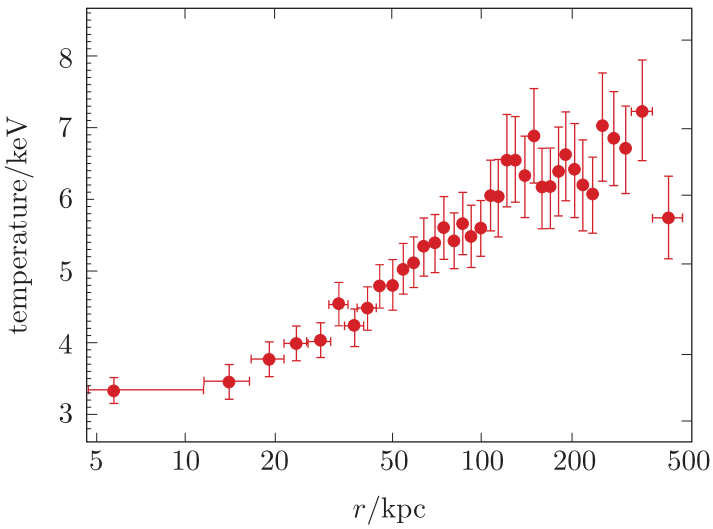
The following example explores typical cooling times in galaxy clusters.

Example 3.3

Figure 3.12 shows the gas density and temperature profiles for Abell 1795 (based on analysis by Cavagnolo *et al.*, 2009). The X-ray luminosity of the inner region ($R < 20\,\mathrm{kpc}$) is $L = 1.0 \times 10^{37}\,\mathrm{W}$.



(a)



(b)

Figure 3.12 (a) Gas number density and (b) temperature as a function of radial distance from the cluster centre for Abell 1795.

By estimating the average temperature and number density in the inner region ($r < 20\,\text{kpc}$), estimate the cooling time for this region. Assume electrons make up 50% of the particles.

If Abell 1795 has existed for most of the age of the Universe, comment on whether a cooling flow is expected.

Solution

The number density and temperature can be roughly estimated from each plot by reading off values at the midpoint radius for the region being considered, i.e. at $\sim 10\,\text{kpc}$.

The plotted density is the electron number density in units of cm^{-3}, and the mean value in the inner 20 kpc is approximately $0.050\,\text{cm}^{-3}$. More precisely, it is a little below this value at 10 kpc, but it is difficult to make a more accurate estimate with the logarithmic axis labels as given here. $1\,\text{cm}^{-3} = 10^6\,\text{m}^{-3}$, and so the estimated mean value corresponds to $50\,000\,\text{m}^{-3}$. Assuming electrons make up 50% of the particles, then the total number density is $100\,000\,\text{m}^{-3}$.

The temperature plot has units that might seem a little unexpected. X-ray astronomers often pre-multiply their temperatures by k_B and work in units of keV because that gives a quantity representing temperature that avoids large exponents. However, the result is technically an energy, not a temperature!

Reading from the plot, an estimate of $k_\text{B}T$ at $\sim 10\,\text{kpc}$ is $\sim 3.4\,\text{keV}$. Converting to SI units gives $5.4 \times 10^{-16}\,\text{J}$. If you are interested to know the value in the more usual units of kelvin (K) you can divide by k_B.

We now need to obtain the cooling function, Λ_cool. We are considering a sphere of radius 20 kpc, and so $V = (4/3)\pi\,(20\,\text{kpc})^3 = 9.8 \times 10^{62}\,\text{m}^3 \approx 10^{63}\,\text{m}^3$, using the appropriate conversion. Therefore, $\Lambda_\text{cool} = L/V \approx 1.0 \times 10^{-26}\,\text{W m}^{-3}$.

We now have all of the information needed to calculate t_cool using Equation 3.15.

$$t_\text{cool} \approx \frac{3 \times \left(100\,000\,\text{m}^{-3}\right) \times \left(5.4 \times 10^{-16}\,\text{J}\right)}{2 \times 1.0 \times 10^{-26}\,\text{W m}^{-3}} = 8.1 \times 10^{15}\,\text{s}$$

Converting to years, we find that the cooling time in the central region of Abell 1795 is $\sim 2.6 \times 10^8$ years, or $\sim 0.3\,\text{Gy}$. Since the age of the Universe is of order 14 Gy, we would expect a cooling flow to have developed in this cluster. This is indeed what X-ray observations show for Abell 1795.

Figure 3.13 shows a set of cooling time profiles for a representative sample of nearby galaxy clusters measured from *Chandra* observations. You will see that in general, t_cool drops below the age of the Universe in the inner 10–100 kpc, with very short cooling times at the centre.

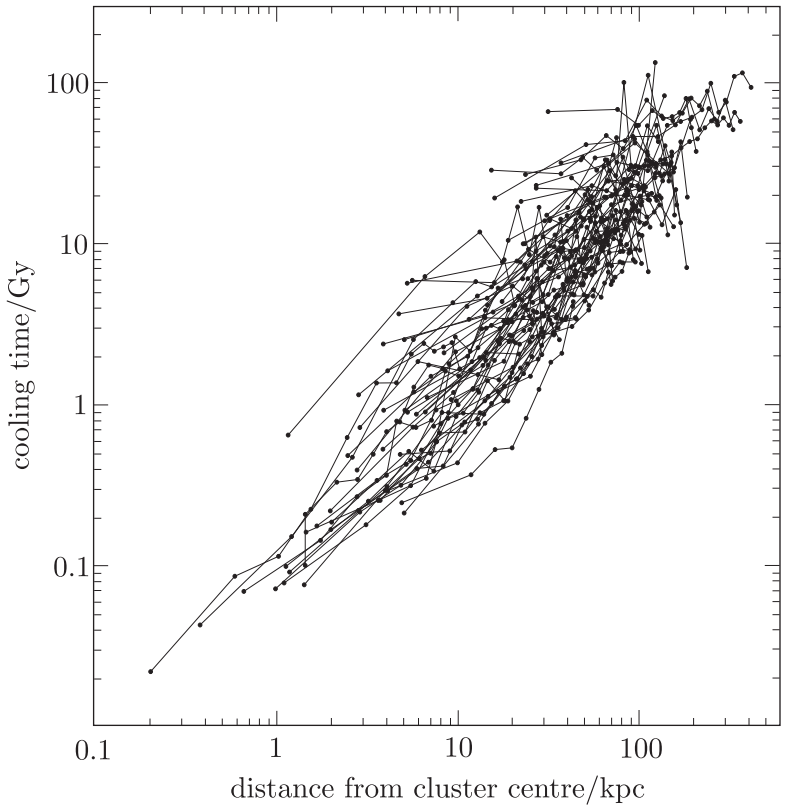


Figure 3.13 A sample of cooling time profiles for nearby galaxy clusters.

Early investigations of cooling flows suggested that the amount of cold gas expected to be present in the central galaxy of a cluster as a result of this process would be extremely large. The rate at which gas should become sufficiently cold (and dense) to form stars in the central galaxy is known as the **mass deposition rate**, $\dot{M}_{\rm cool}$, calculated as

$$\dot{M}_{\rm cool} = \frac{M_{\rm gas}(r < r_{\rm cool})}{t_{\rm cool}} \tag{3.16}$$

where $r_{\rm cool}$ is typically defined as the radius within which the cooling time is less than the Hubble time (so that cooling can be significant over the cluster's lifetime).

For a cluster such as Abell 1795, $r_{\rm cool}$ is of order tens of kpc and the expected rate of deposited cold gas into the central galaxy would be $\dot{M}_{\rm cool} \sim 100\,{\rm M}_\odot\,{\rm y}^{-1}$. It is possible to convert this into a rate of predicted star formation, assuming a typical star formation efficiency (e.g. estimated from studies of the Milky Way).

Much observational effort has been devoted to testing whether the high star-formation rates and cold gas in central galaxies that such calculations imply is indeed present and affecting central galaxy evolution. Figure 3.14 shows that – for $\dot{M}_{\rm cool} \gtrsim 20\,{\rm M}_\odot\,{\rm y}^{-1}$ – star-formation rate does depend on the predicted rate of mass deposition calculated from Equation 3.16.

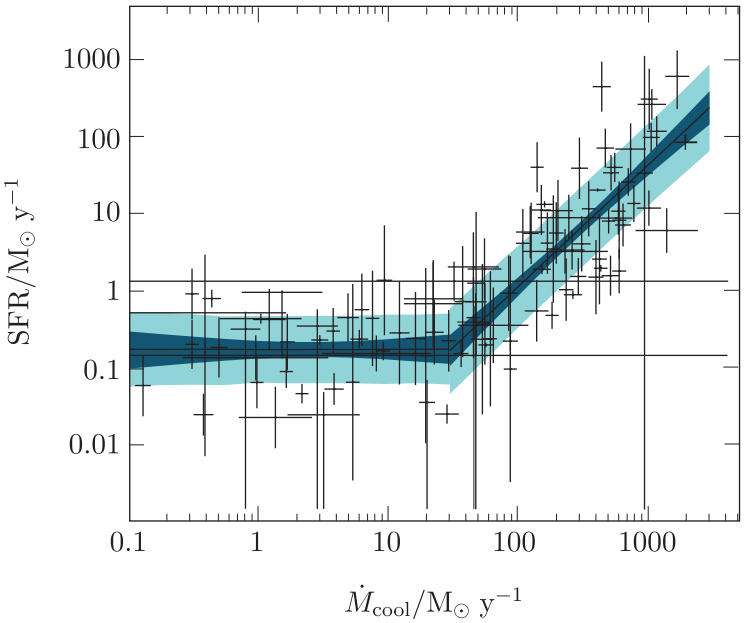


Figure 3.14 A comparison of star-formation rate in a sample of BCGs with the mass deposition rates inferred from their X-ray luminosity. The dark turquoise region represents the best model fits to the data, while the lighter regions show a measure of the scatter in the data.

The next exercise compares the star formation of BCGs shown in Figure 3.14 with that of the Milky Way.

Exercise 3.5

The average star-formation rate of the Milky Way over its recent history is $\sim 2\,\mathrm{M}_\odot\,\mathrm{y}^{-1}$, while the rate of gas falling into the Milky Way from its wider environment (the Local Group) is $< 1\,\mathrm{M}_\odot\,\mathrm{y}^{-1}$. Consider a typical bright galaxy cluster whose mass deposition rate, inferred from its X-ray luminosity, is $200\,\mathrm{M}_\odot\,\mathrm{y}^{-1}$. Use Figure 3.14 to infer the expected star-formation rate in the cluster's central galaxy. What is different about the behaviour of this BCG and the Milky Way, and how might this affect their subsequent evolution?

The previous exercise shows that the evolution of a cluster-centre galaxy is strongly affected by the behaviour of the ICM. If the X-ray inferred $\dot{M}_{\rm cool}$ values are correct, then a BCG should contain an increasing and very large reservoir of cold gas.

But it is important to emphasise that $\dot{M}_{\rm cool}$ is the inferred rate of mass deposition *only if* the energy lost to X-ray radiation is not in some way replenished. In fact, although BCGs do contain some molecular gas, they do not typically contain enough to match the expected deposition rates.

There is now considerable evidence that the lost energy *is* at least partially replenished via the processes of galaxy feedback, which is explained in the next section.

3.2.4 Galaxy feedback in clusters

Many galaxies in rich environments, and especially BCGs, host active galactic nuclei (AGN), in which accretion of material onto the central supermassive black hole leads to high-energy radiation and outflows of gas. The energetic processes associated with AGN – particularly with the jets of **radio galaxies** – transport energy outwards from galaxy centres, which can heat up the intracluster gas and compensate for the energy lost to X-ray radiation discussed in the previous section. In Chapter 4 you will consider the physics of AGN jets in more detail, but here we focus on how they can affect galaxies in clusters.

Figure 3.15 shows an example of this energy transport in action: radio-emitting bubbles produced by the AGN are shown in red, embedded in the intracluster medium (blue), and extend for a large distance beyond the central galaxy that is the origin of the outflowing material.

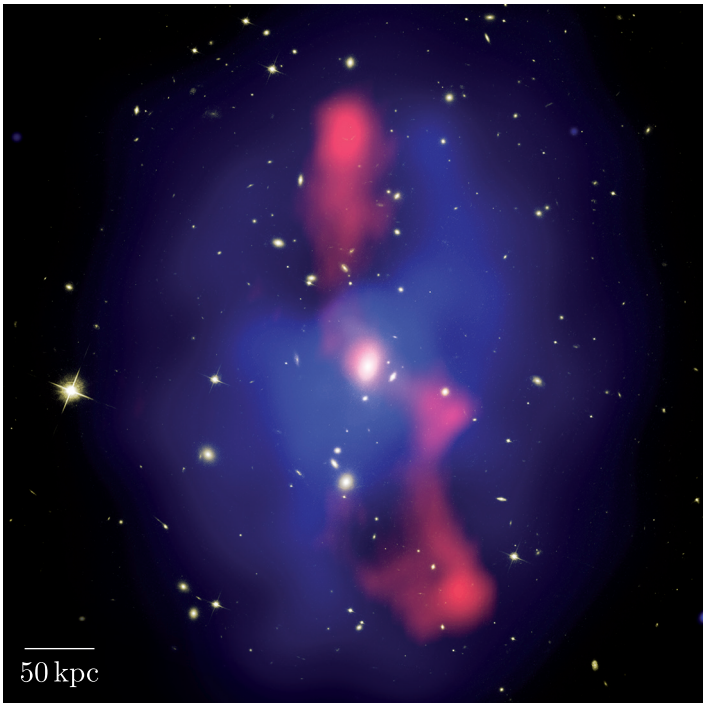


Figure 3.15 A BCG contains an AGN that produces jet-driven outflows (red regions in this image) extending to a large distance.

There is strong observational evidence that radio galaxies, like the example shown in Figure 3.15, distribute a lot of energy away from the central region of the BCG and into the surrounding ICM. It is thought that a **galaxy feedback** cycle is in operation, as shown in Figure 3.16: gas cooling from the ICM falls into the central galaxy, feeding the central black hole and driving the radio jets and bubbles. In turn this outflow heats the surrounding gas, which limits the amount of cooling that can take place. Clusters could either go through repeated cycles of cooling and heating, or remain in a state where the two processes are roughly in balance.

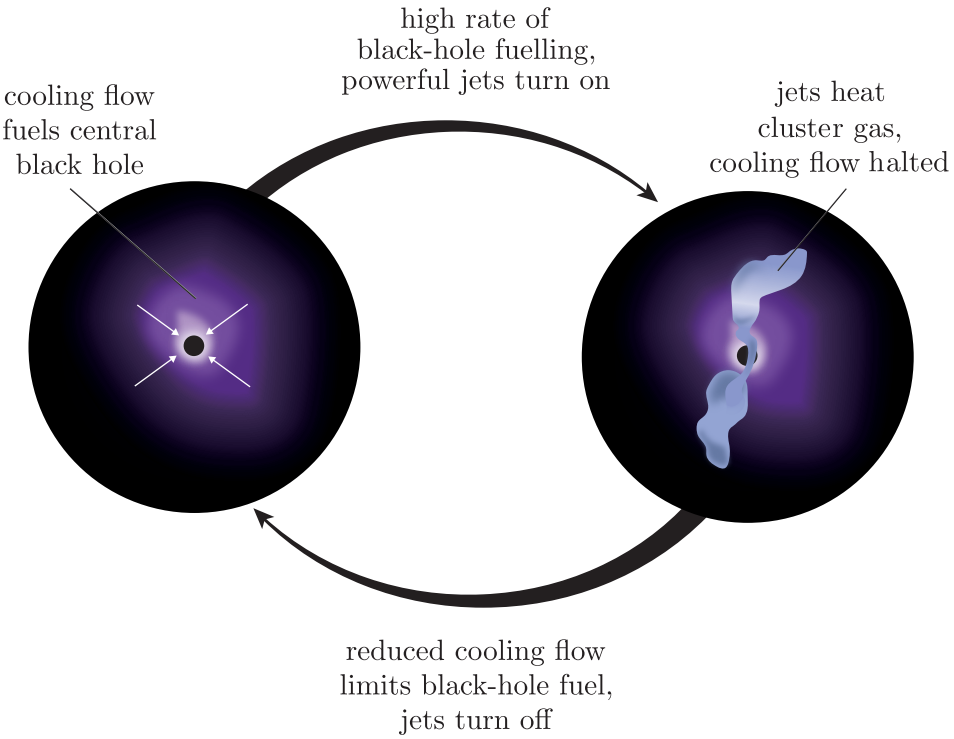


Figure 3.16 The cycle of AGN feedback in galaxy clusters.

There is a great deal of observational evidence that such a feedback cycle operates in nearby galaxy clusters. One compelling form of evidence comes from comparing estimates of the typical lifetimes of radio galaxies with cluster cooling times ($t_{\rm cool}$). It is possible to estimate the age of a radio galaxy using $t = d/v$, where d is the length from the nucleus to the outer edge (i.e. the maximum distance travelled by the outflow), and v is an assumed expansion speed. The slowest likely expansion speed is the local **sound speed** in the ICM, given by

$$c_{\rm s} = \sqrt{\frac{5k_{\rm B}T}{3\langle m\rangle}} \tag{3.17}$$

where $\langle m\rangle = 0.6m_{\rm p}$ is the mean particle mass.

Exercise 3.6

Make an estimate of the age of a radio galaxy whose emission extends to a radius of 200 kpc within a galaxy cluster of temperature 5×10^7 K. Comment on how this would change if the radio-galaxy expansion speed is supersonic by a factor of 2 or 3. Compare your estimates to typical cooling times in cluster centres.

As well as considering individual clusters, we can also look at the whole population of clusters, and investigate whether all of the AGN activity in the local Universe can provide enough heating to compensate for the energy lost to X-ray radiation.

In *Cosmology* Chapter 11 you were introduced to luminosity functions, which describe how the number density of galaxies depends on their luminosity. The following example uses the luminosity function of galaxy clusters to make a rough estimate of the total rate of energy loss that would need to be compensated for by heating.

Example 3.4

Assume that the number density of galaxy clusters, n_{clus}, within the range of X-ray luminosities L_{X} to $L_{\text{X}} + \mathrm{d}L_{\text{X}}$ is given by

$$n_{\text{clus}}(L_{\text{X}})\,\mathrm{d}L_{\text{X}} = \frac{n_0}{L_{\text{X}^*}}\left(\frac{L_{\text{X}}}{L_{\text{X}^*}}\right)^{-\alpha}\mathrm{d}L_{\text{X}} \tag{3.18}$$

where $\alpha = 1.8$, $n_0 = 4.5 \times 10^{-7}\,\text{Mpc}^{-3}$ and $L_{\text{X}^*} = 3.0 \times 10^{37}\,\text{W}$ are constants (the slope and normalising factors for number density and luminosity).

Calculate the total rate of energy loss via X-ray radiation per cubic megaparsec, assuming a typical X-ray luminosity range of 10^{35}–$10^{38}\,\text{W}$.

Solution

To sum up the luminosity produced by all of the galaxy clusters in a given volume, it is necessary to perform an integral over Equation 3.18.

If we simply took the integral of the expression provided, this would give us the total number density of galaxy clusters falling within the given luminosity range. The luminosity density (total luminosity per unit volume), ϵ_{L}, for a particular L_{X} range is given by multiplying Equation 3.18 by L_{X}:

$$\begin{aligned}\epsilon_{\text{L}}(L_{\text{X}})\,\mathrm{d}L_{\text{X}} &= n_0 \frac{L_{\text{X}}}{L_{\text{X}^*}}\left(\frac{L_{\text{X}}}{L_{\text{X}^*}}\right)^{-\alpha}\mathrm{d}L_{\text{X}} \\ &= n_0\left(\frac{L_{\text{X}}}{L_{\text{X}^*}}\right)^{1-\alpha}\mathrm{d}L_{\text{X}}\end{aligned}$$

We can now sum up the luminosity density over the full luminosity range by integrating this expression over the given range of X-ray luminosities:

$$\epsilon_{\text{tot}} = \int_{L_1}^{L_2} n_0\left(\frac{L_{\text{X}}}{L_{\text{X}^*}}\right)^{1-\alpha}\mathrm{d}L_{\text{X}}$$

where L_1 and L_2 correspond to the X-ray luminosity range given in the question.

Taking the constants outside the integral gives

$$\epsilon_{\text{tot}} = \frac{n_0}{L_{\text{X}^*}^{1-\alpha}}\int_{L_1}^{L_2} L_{\text{X}}^{1-\alpha}\,\mathrm{d}L_{\text{X}}$$

which becomes

$$\epsilon_{\text{tot}} = \frac{n_0}{L_{\text{X}^*}^{1-\alpha}(2-\alpha)}\left[L_{\text{X}}^{2-\alpha}\right]_{L_1}^{L_2}$$

We can now substitute in the given values of n_0, L_{X^*}, α, and the upper and lower X-ray luminosity limits of $L_1 = 10^{35}\,\mathrm{W}$ and $L_2 = 10^{38}\,\mathrm{W}$, which gives a total luminosity density of $\epsilon_{\mathrm{tot}} = 6.4 \times 10^{31}\,\mathrm{W\,Mpc^{-3}}$.

Example 3.4 leads to an estimated rate at which clusters lose energy of $\sim 6 \times 10^{31}\,\mathrm{W\,Mpc^{-3}}$. This can be compared to the possible rate at which AGN could reheat the gas to replenish the energy loss. The rate of heating potentially available from a pair of AGN jets can be estimated as equivalent to the mechanical power of the jets, i.e. the energy per unit time travelling up the jets, usually given the symbol Q (to avoid confusion with pressure, P). The next exercise explores the amount of jet heating available from AGN in the local Universe.

Exercise 3.7

An approximation of the number of radio galaxies, n_{RG}, within a range of jet power, Q to $Q + \mathrm{d}Q$, in the Universe out to $z \approx 0.5$ is given by

$$n_{\mathrm{RG}}(Q)\,\mathrm{d}Q = \frac{n_0}{Q_*}\left(\frac{Q}{Q_*}\right)^{-\beta}\mathrm{d}Q$$

where the slope $\beta = 1.65$, the number density normalisation $n_0 = 1.3 \times 10^{-6}\,\mathrm{Mpc^{-3}}$ and the jet power normalisation $Q_* = 10^{37}\,\mathrm{W}$.

Using a similar approach to Example 3.4, make an estimate of the total heating rate per unit volume, ϵ_{RG}, from radio galaxies in the range of jet power between $10^{35}\,\mathrm{W}$ and $10^{38}\,\mathrm{W}$.

Compare your heating rate estimate to the rate of energy loss from clusters estimated in Example 3.4.

As suggested by the various (fairly rough) comparisons in this section, it is now thought that heating greatly reduces the amount of gas that cools and forms stars in BCGs. The biggest and most massive galaxies in the Universe have their growth halted by the feedback from AGN jets.

3.3 Summary of Chapter 3

- Galaxy clusters can be characterised by their **optical richness**, which is a measure of the number of galaxies they contain and scales with the total cluster mass (dominated by dark matter).
- The **intracluster medium** (ICM) is a low-density gas that pervades the space between galaxies in a cluster.
- The ICM is hot and emits X-rays via the **thermal bremsstrahlung** process, as well as via electronic transitions between energy levels of ions. The rate at which X-rays are produced is given by

$$\epsilon = 1.4 \times 10^{-40} g_{\mathrm{ff}} Z^2 n_{\mathrm{e}} n_{\mathrm{i}} T^{1/2}\,\mathrm{W\,m^{-3}} \qquad \text{(Eqn 3.2)}$$

- The ICM is in a state of **hydrostatic equilibrium**, in which gravitational and pressure forces balance at each radius:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2} \qquad \text{(Eqn 3.7)}$$

- The assumption of hydrostatic equilibrium means that X-ray measured profiles of gas density and temperature can be used to measure the total mass distribution in a cluster, according to

$$M(r) = -\frac{k_\mathrm{B}r^2}{G\langle m\rangle}\left(\frac{\mathrm{d}T}{\mathrm{d}r} + \frac{T(r)}{\rho(r)}\frac{\mathrm{d}\rho}{\mathrm{d}r}\right) \qquad \text{(Eqn 3.8)}$$

- The **Sunyaev–Zeldovich effect** (SZ effect) is a shift in the observed frequency of cosmic microwave background photons that pass through galaxy clusters, caused by inverse Compton scattering. The signal strength is proportional to gas pressure, and is independent of distance, so it can be used to study distant clusters.
- Cluster environment influences how galaxies evolve by increasing the likelihood of galaxy mergers, tidal forces, and processes related to the influence of the ICM.
- Cluster galaxies are more likely to be ellipticals, and isolated galaxies are more likely to be spiral – this is known as the **morphology–density relation**.
- **Ram pressure stripping** can remove gas from galaxies as they travel through the ICM, with the pressure related to the galaxy's speed:

$$P_\mathrm{ram} = \rho v^2 \qquad \text{(Eqn 3.12)}$$

  Gas will be stripped from the galaxy if P_ram exceeds the gravitational force per unit area binding the gas to the galaxy disc:

$$\frac{F_\mathrm{grav}}{A} = 2\pi G\Sigma_*\Sigma_\mathrm{gas} \qquad \text{(Eqn 3.13)}$$

- The high X-ray luminosities of cluster centres lead to energy loss and short cooling times in the central region:

$$t_\mathrm{cool} = \frac{3nk_\mathrm{B}T}{2\Lambda_\mathrm{cool}} \qquad \text{(Eqn 3.15)}$$

- If the energy lost to X-ray radiation isn't replenished then a **cooling flow** will develop and cold gas from the ICM will be deposited onto the central galaxy of the cluster, leading to enhanced rates of star formation in that galaxy.
- It is now known that **galaxy feedback** via powerful jets from active galactic nuclei heats the intracluster medium and balances cooling. This feedback helps to regulate star formation so that the most-massive galaxies grow more slowly than would otherwise be expected.

Chapter 4 Black-hole jets

Black holes are now known to be ubiquitous throughout the Universe, with every galaxy having its own central supermassive black hole with a mass millions to billions of times larger than that of the Sun. The in-fall (accretion) of gas onto black holes in galaxy centres causes energetic outflows in the form of jets and winds, as well as the production of large amounts of radiation across the electromagnetic spectrum. As you saw at the end of the previous chapter, black-hole jets have an important influence on how massive galaxies evolve.

This chapter will focus on the physics of black-hole jets. Jets are produced by both stellar-mass and supermassive black holes, but in this chapter we will be considering only extragalactic jets from SMBHs in distant galaxies. As with previous chapters you will explore both small-scale processes that influence how we observe jets, and large-scale processes that transport energy within and beyond the galaxy environments of SMBHs.

Objectives

Working through this chapter will enable you to:

- describe the observed properties of radio galaxies and radio-loud quasars
- summarise the evidence for relativistic bulk speeds for black-hole jets
- explain how special relativity affects observations of jets
- describe and solve problems relating to the process of synchrotron radiation
- solve problems related to the total internal energy of radio galaxies, the nature of how jets are powered and the energy available for galaxy feedback.

4.1 Observing black-hole jets

The first identified jet originating from a black hole was a ‘curious straight ray’ seen in optical images of the central galaxy in the Virgo cluster, Messier 87 (M87) by the astronomer Heber Curtis in 1918, over a decade before Hubble’s discovery of the expansion of the Universe. The advent of radio astronomy after the Second World War led to the discovery of large numbers of jets emitting at radio wavelengths. The outflowing material in M87 has now been studied by many facilities including, most recently, the Event Horizon Telescope, which imaged its black-hole shadow. Figure 4.1 shows a montage of radio observations of the M87 jet, revealing its highly complex structure on many different size scales, extending far beyond the optical galaxy itself.

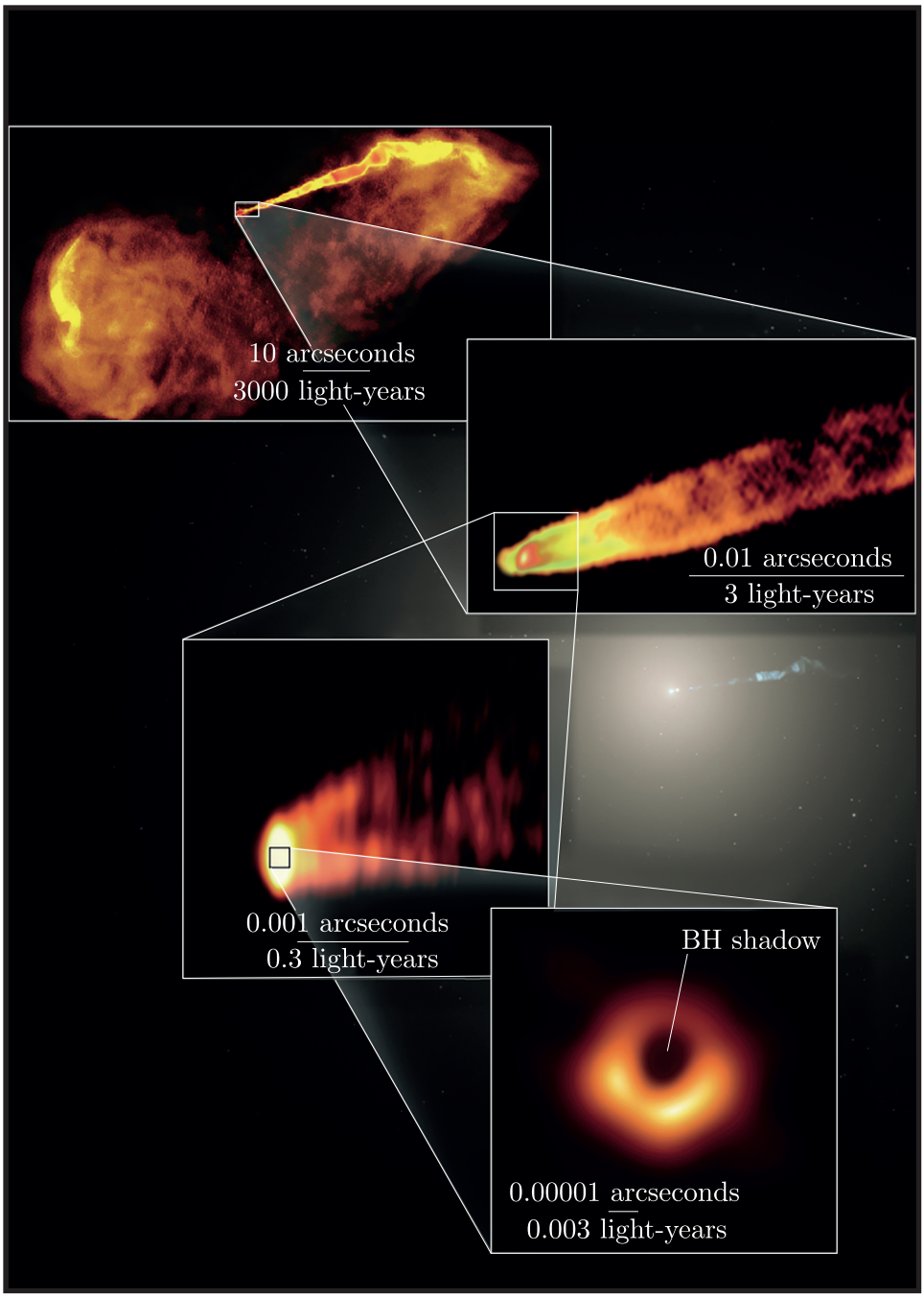


Figure 4.1 Four observations of the M87 radio jet from much less than a parsec to kiloparsec scales, as seen at various radio frequencies by the Very Large Array (VLA), Very Long Baseline Array (VLBA), Global 3 mm VLBI Array (GMVA) and Event Horizon Telescope (EHT).

M87 is a fascinating example of black-hole behaviour because, as one of the nearest examples of a black-hole jet, we can observe not only the currently active jet, but also fainter radio emission on much larger scales, which originates from plasma transported up the jet over very long timescales. Figure 4.2 compares this larger-scale radio structure – which is

spread throughout the central regions of the Virgo cluster – to the VLA image of the 3000 light-year jet shown in Figure 4.1.

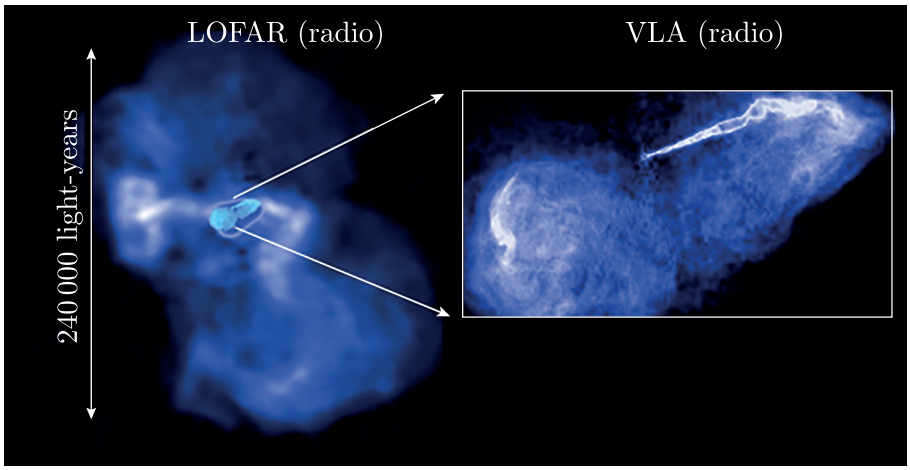


Figure 4.2 Low-Frequency Array (LOFAR) image of the plasma from the M87 jet on galaxy-cluster scale, compared with the 3000 light-year jet structure shown in Figure 4.1.

Galaxies possessing large-scale radio-emitting jets are known as **radio galaxies** and **radio-loud quasars** (depending on whether they also possess the optical characteristics of quasars). These two classes form part of the population known as **active galaxies**, which are systems where the electromagnetic radiation in some part of the spectrum outshines the starlight from the galaxy (see Chapter 1).

The difference between radio galaxies and radio-loud quasars relates to the properties of the central active galactic nucleus (AGN). Quasars, which can be either radio-loud or radio-quiet (not possessing large-scale jets), have an optically bright central region (nucleus) whereas radio galaxies do not. This is related to the in-fall of material onto the central black hole, and is also affected by the galaxy's orientation relative to us. We return to the influence of orientation on the brightness of AGN features in later sections.

Modern radio surveys reveal that radio-emitting jets are present in millions of galaxies, and that all massive galaxies are likely to have gone through a phase of jet production. In the next sections you will learn about the physics of these jets. In particular, we will discuss the evidence that the outflowing material travels at close to the speed of light, and the reasons why it has long been accepted that the central engine producing the jets must be a supermassive black hole.

4.1.1 Evidence for relativistic outflows

One of the reasons why radio jets are of interest is that they are the most easily studied example of a **relativistic outflow**: there are multiple lines of convincing evidence that the typical speed of matter travelling outward along the jets is close to the speed of light.

The simplest evidence comes from directly observing the movement of bright blobs of jet emission over time, and so inferring their speed. Example 4.1 considers in what situations such direct measurements are possible.

Example 4.1

Consider three radio galaxies, M87 ($d_A = 16\,\text{Mpc}$), Centaurus A ($d_A = 3.4\,\text{Mpc}$) and Cygnus A ($d_A = 206\,\text{Mpc}$), where d_A is the angular diameter distance to each galaxy. For each galaxy consider three possible jet speeds of (i) $1000\,\text{km s}^{-1}$, (ii) $0.01c$ and (iii) $0.9c$, and for each speed determine the angular distance in units of arcseconds that a blob of jet material could travel in between two monitoring observations taken 5 years apart. For simplicity, you should assume that the jet is oriented in the plane of the sky (i.e. perpendicular to our line of sight to the galaxy, such that all locations along the jet are the same distance from us).

Solution

For each radio galaxy, we need to determine the angular distance on the sky that corresponds to the distance a blob of material will travel at each speed in a time interval of 5 years.

We can first calculate the physical distance travelled for each speed, which is the same for all of the galaxies. Using $l = vt$, we obtain distances of (i) $1.58 \times 10^{14}\,\text{m}$, (ii) $4.73 \times 10^{14}\,\text{m}$ and (iii) $4.26 \times 10^{16}\,\text{m}$.

To convert these physical distances to angles on the sky, we use the simple geometric relationship $\theta = l/d_A$, where θ is measured in radians, l is the physical distance travelled and d_A is the distance to the galaxy.

Applying this relationship, and converting the resulting angles from radians to arcseconds, gives the results in Table 4.1.

Table 4.1 Angular distances (in arcseconds) that a blob of jet material moves on the sky over a 5-year period.

Jet speed	M87	Cen A	Cygnus A
(i) $1000\,\text{km s}^{-1}$	6.6×10^{-5}	3.1×10^{-4}	5.1×10^{-6}
(ii) $0.01c$	2.0×10^{-4}	9.3×10^{-4}	1.5×10^{-5}
(iii) $0.9c$	0.018	0.084	0.0014

The angles on the sky that jet material can travel on a 5-year timescale are therefore very small: only fractions of an arcsecond.

The previous example demonstrates that observations of very high angular resolution are needed to measure the motion of jets directly. The most widely used radio telescopes have the capability of distinguishing features on scales of <1 arcseconds, with some specialist instruments able to see details in the brightest jets of $\sim 10^{-3}$ arcseconds.

Another way of thinking about it is that *if* motion is detected via long-term observations, this demonstrates that jets must be travelling at very high speeds. Figure 4.3 shows the result of radio monitoring observations of Centaurus A over an 11-year period in which several jet knots (compact regions of bright emission) have moved by a very small, but detectable, angle on the sky. Several regions of the jet are found to be moving at apparent speeds of $\sim 0.5c$.

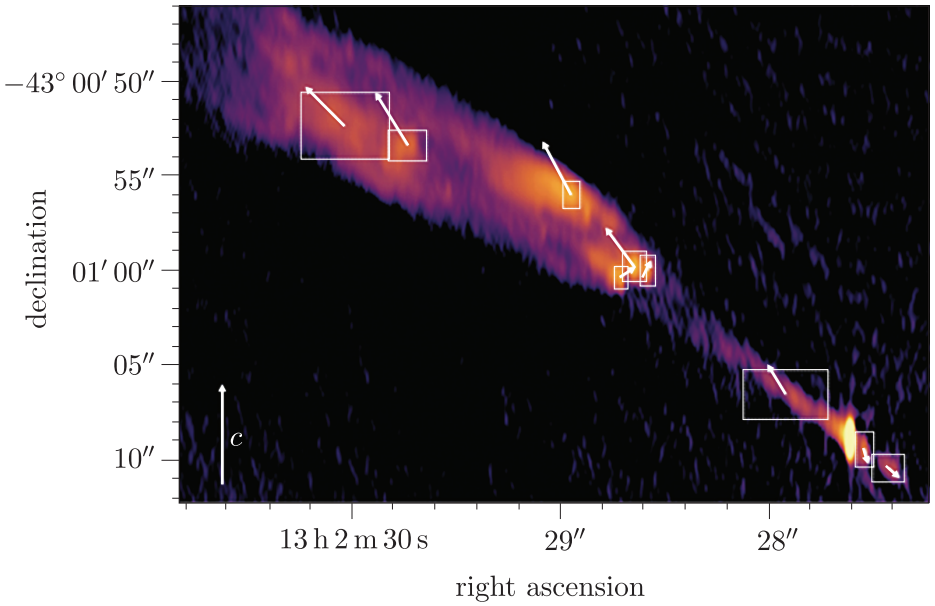


Figure 4.3 The results of monitoring the Centaurus A radio jet over a period of 11 years (after Hardcastle *et al.*, 2003). Arrows show the direction and apparent speed of motion for each moving knot or region of jet fluid. Arrow lengths are scaled to indicate the speed relative to c, which is the vertical arrow in the bottom left-hand corner.

- If jet material travels at half of the speed of light, would you expect special relativity effects, such as time dilation, to be relevant to observations of jets?
- At $v = 0.5c$, the Lorentz factor, $\gamma = 1/\sqrt{1-(v/c)^2} = 1.15$. Since this is larger than 1 by a non-trivial amount (15%), effects such as time dilation are important for interpreting observations of this jet.

The situation gets even more interesting when some of the brightest known jets are observed at high angular resolution. Figure 4.4 shows an example of the phenomenon known as **apparent superluminal motion** – the right-hand blob at the end of the jet emission appears to travel a distance of over 25 light-years in a little over six years of observations.

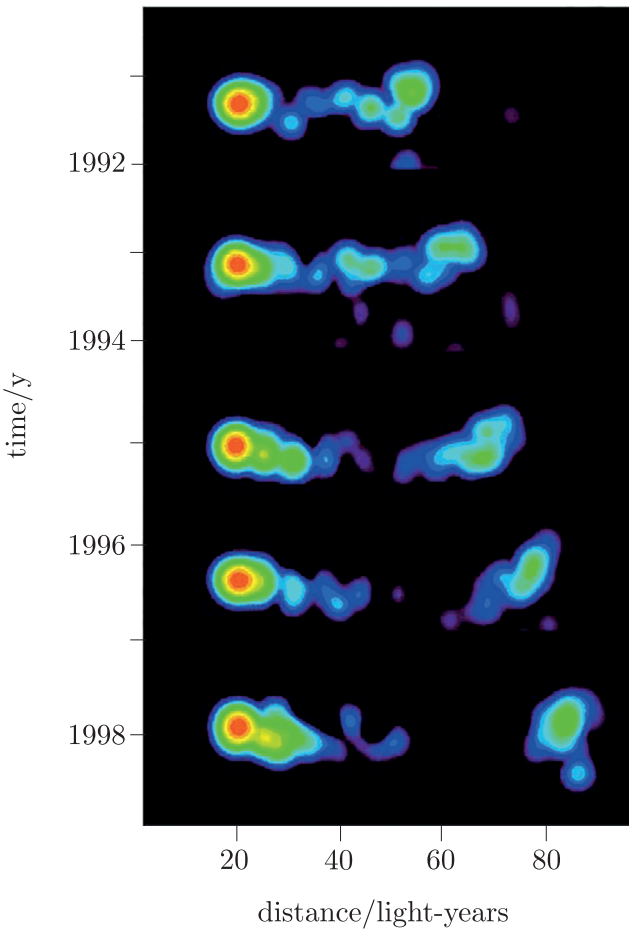


Figure 4.4 The apparent superluminal motion of a radio jet over time. The red region corresponds to a stationary radio-bright region near to the AGN, with the blobs to the right moving outwards with time.

What is going on here? If material cannot travel faster than the speed of light, then this must be some sort of optical illusion. The discussion above might lead you to think that time dilation or similar effects may be the cause. In fact the explanation is simpler. We have neglected to consider the geometry of the situation, and the distance inferred in the figure does not correct for the fact that the jet may not be oriented in the plane of the sky. This means that if the material in the jet is travelling at relativistic speeds then its motion in the direction towards us is nearly as fast as that of the light that the material is emitting.

Example 4.2 considers how geometry affects the speeds we infer for observed jets, such as the example in Figure 4.4.

Example 4.2

Consider a jet that is oriented at an angle, θ, towards the line of sight to the Earth, as indicated by the diagram shown in Figure 4.5. A blob of jet material is observed to move from position A at time t_1 to position B at time t_2 (where times are in the rest frame of the observer). Use basic geometry to derive a relationship between the measured speed of the jet in the direction perpendicular to the observer's line of sight, $v_{\rm app}$, and the true speed of the jet in the direction it is travelling, V.

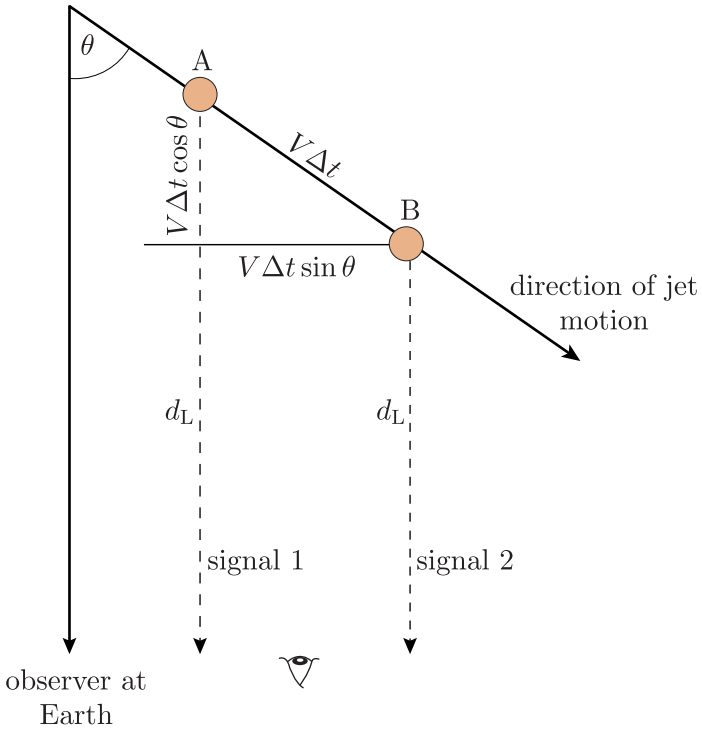


Figure 4.5 The geometry of the jet considered in Example 4.2. Note that $d_{\rm L}$ is very large compared to the other distances in the image.

Solution

The true distance travelled by the jet blob in time interval $\Delta t = t_2 - t_1$ is $V\Delta t$. The apparent speed, $v_{\rm app}$, measured at Earth (as, for example, inferred from Figure 4.4) is given by the transverse distance travelled by the jet blob in the direction perpendicular to the observer's line of sight in a time interval $\Delta t_{\rm obs}$, which is the interval between the arrival of light signals emitted at time t_1 (when the blob was at position A) and the light emitted at time t_2 (when the blob was located at B).

We can consider the arrival times of each of these two signals, $t_{1,\rm obs}$ and $t_{2,\rm obs}$, using $t = d/v$ where $v = c$ for the light signals, and d is the distance from each location to the observer at Earth.

If we define d_L as the distance from the nearer location, B, to the observer, as shown in Figure 4.5, then

$$t_{2,obs} = t_2 + d_L/c$$

and

$$t_{1,obs} = t_1 + [d_L + V\Delta t \cos\theta]/c$$

We can now evaluate the apparent jet speed measured by the observer at Earth, which is the transverse distance, divided by Δt_{obs}:

$$v_{app} = \frac{V\Delta t \sin\theta}{\Delta t_{obs}} \tag{4.1}$$

The observed interval between signals from the two locations, $\Delta t_{obs} = t_{2,obs} - t_{1,obs}$, so

$$\Delta t_{obs} = t_2 + d_L/c - t_1 - d_L/c - (V/c)\,\Delta t \cos\theta$$

which simplifies to

$$\Delta t_{obs} = \Delta t[1 - (V/c)\cos\theta]$$

Therefore Equation 4.1 can be expanded as

$$v_{app} = \frac{V\Delta t \sin\theta}{\Delta t[1 - (V/c)\cos\theta]}$$

which simplifies to

$$v_{app} = \frac{V \sin\theta}{1 - (V/c)\cos\theta} \tag{4.2}$$

It is apparent from Equation 4.2 that if $V/c \ll 1$ (i.e. the jet speed is not relativistic) then the apparent jet speed cannot be greater than the true jet speed, because the denominator reduces to 1, and $\sin\theta$ is always <1. However, the situation is more interesting when V approaches c.

Equation 4.2 is also commonly written in terms of $\beta = V/c$, as follows:

$$\beta_{app} = \frac{\beta \sin\theta}{1 - \beta\cos\theta} \tag{4.3}$$

where $\beta_{app} = v_{app}/c$.

Exercise 4.1 gives you the chance to apply the results of Example 4.2 and so investigate how Equation 4.3 provides an explanation for jet velocity measurements that appear superluminal.

Exercise 4.1

Calculate the apparent jet speed, β_{app}, that would be measured for all combinations of jet angle $\theta = [1°, 10°, 25°]$ and (true) jet speed $\beta = [0.5, 0.9, 0.99]$.

(*Hint*: writing a short Python code may make this calculation easier.)

Exercise 4.1 shows that apparent superluminal motion can be observed for a wide range of jet speeds and source angles, provided V is an appreciable fraction of c. The fact that apparent superluminal motion can only occur for relativistic speeds provides one compelling form of evidence that jets are relativistic.

4.1.2 Relativistic beaming

Now that we have some evidence that jets are relativistic we can begin to consider how this affects other jet measurements we might make. The Lorentz transformations (*Cosmology* Chapter 2) affect the relationship between the quantities we measure in observations and the *intrinsic* properties of the jet (i.e. those that would be measured by an observer travelling with the jet). For example, if we don't account for the Lorentz transformations then we will make incorrect measurements of the true length or volume of a jet, with knock-on effects if, for example, we want to consider how much energy the jet transports.

One of the most important effects of relativity is a phenomenon known as **relativistic boosting**. The visible impact of this effect is that electromagnetic radiation from material flowing towards us at a relativistic speed is brighter than would otherwise be expected. Conversely, emission from material travelling away from us appears dimmer.

The radio galaxy M87 is a good example of this phenomenon. In the top panel of Figure 4.1 the largest scale emission shows clear evidence that an outflow is present on both sides of the nucleus, but the bright central jet is only visible in one direction from the bright core. The asymmetry of observed jets (despite other evidence that energy is actually being transported in both directions) provides strong evidence that jets have relativistic speeds, for reasons we will now explore further.

There are several effects that together contribute to the boosting of emission from jet material travelling in a direction angled towards the observer. Figure 4.6 shows the first effect, known as **aberration**, in an everyday context involving non-relativistic speeds. Note that panel (b) shows the rain as viewed in a reference frame moving with the observer who is shown.

Aberration is the change in angle of moving objects (such as raindrops) as seen from the perspective of observers moving at different speeds. The rain appears to be angled towards a moving person, because they are effectively catching up with the initially more distant raindrops. This makes it necessary to hold the umbrella at an angle in order to stay dry.

A second important effect is that, for a moving observer, the *rate* at which the rain hits the umbrella changes (this is why rain seems heavier if you are driving on a motorway). This phenomenon is related to the **Doppler effect**, in which the frequency of waves (such as the sound waves produced by an ambulance zooming past) is changed by relative motion.

(a)

(b)

Figure 4.6 The effect of aberration of raindrops: a stationary observer (a) sees the rain falling vertically from the sky, whereas from the perspective of a moving observer (b) the rain will be falling at an angle.

Both aberration and the Doppler effect are important in non-relativistic situations. But both effects involve intervals in time and space, and so we need to use special relativity to account for their influence on measurements of relativistic jets.

To investigate these relationships, we need to make use of the Lorentz transformations for velocity, which are listed in the following box (in the form in which the primed quantities are known and we wish to calculate the unprimed velocities). Here we assume the standard configuration of two inertial reference frames, S (our frame) and S′, where the latter is the frame moving with the jet so it's travelling at a speed of V in the positive x-direction.

Lorentz transformations for velocity

$$v_x = \frac{v'_x + V}{1 + Vv'_x/c^2} \tag{4.4}$$

$$v_y = \frac{v'_y}{\gamma(1 + Vv'_x/c^2)} \tag{4.5}$$

$$v_z = \frac{v'_z}{\gamma(1 + Vv'_x/c^2)} \tag{4.6}$$

- If the two frames are defined by relative motion in the x-direction, then why isn't $v_x = V$?
- Equations 4.4, 4.5 and 4.6 describe the situation where we want to measure the velocity of an object that is moving in an arbitrary direction in frame S, and so v_x refers to the x-component of the motion of this object (which is unrelated to the relative motion of the two frames).

Comparing Equations 4.5 and 4.6 with the Lorentz transformations for the *y- and z-positions* (*Cosmology* Chapter 2) reveals an important difference that is linked to the idea of aberration. For relative motion in the x-direction, the y- and z-coordinate positions of an object *do not change*, but the components of the object's velocity in those directions *do change*.

The components of motion in different directions become intertwined. This has the important consequence that the net angle of the motion we observe changes, as Example 4.3 demonstrates.

Example 4.3

Consider a photon that is emitted by a region of jet that is moving at a relativistic speed V in our (observer's) frame of reference S. As shown in Figure 4.7, the photon travels at an angle θ', defined relative to the x'-axis, which is in the standard configuration in the direction of the jet's motion.

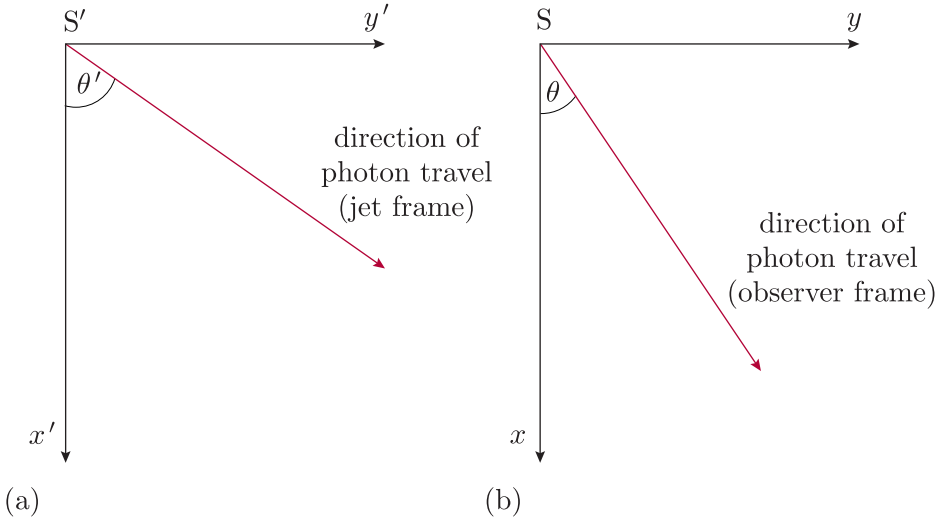


Figure 4.7 Diagram of photon travel direction as measured in (a) the jet frame S′ and (b) the observer frame S.

Recalling that $\tan\theta = v_y/v_x$ (where the velocity components refer to the photon), use the Lorentz velocity transformations to derive an expression for $\tan\theta$, the angle at which we observe the photon to travel in our frame S.

Solution

We are aiming to derive an expression for $\tan\theta$ and so we can start by expressing it in terms of v_x and v_y:

$$\tan\theta = \frac{v_y}{v_x}$$

We can now use the Lorentz transformations for velocity to substitute in for v_x and v_y

$$\tan\theta = \frac{v_y}{v_x} = \frac{v_y'}{\gamma(v_x' + V)}$$

where we have cancelled out the terms $(1 + Vv_x'/c^2)$ from the denominators of both velocity components.

Since we are considering the velocity of a photon, the magnitude of its velocity vector is $v' = c$, and so we can write the x- and y-components as $v_x' = c\cos\theta'$ and $v_y' = c\sin\theta'$, and so

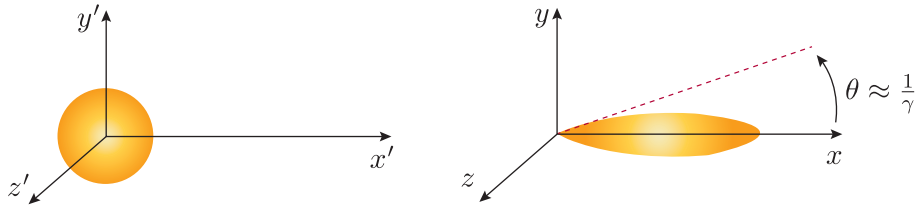
$$\tan\theta = \frac{1}{\gamma}\frac{c\sin\theta'}{c\cos\theta' + V} = \frac{1}{\gamma}\frac{\sin\theta'}{\cos\theta' + V/c} \tag{4.7}$$

We have therefore shown that the observed angle will depend both on the emitted angle of the photon relative to the jet direction and on the jet speed (V). In the next exercise you can examine what this means for photons emitted by a jet.

Exercise 4.2

Consider a photon emitted by a region of jet material that is travelling at speeds relative to the Earth of (i) $0.5c$, (ii) $0.95c$ and (iii) $0.99c$. In each case the photon is emitted perpendicular to the jet direction, namely $\theta' = \pi/2$ radians (or 90°). For each speed, calculate the angle, θ, corresponding to the photon's direction that we observe in our frame of reference S, observing from the Earth.

The preceding exercise demonstrates that as jet speeds get close to c, the aberration effect beams any emitted radiation into a very narrow cone of angles along the direction of motion of the jet. This **relativistic beaming** effect is illustrated in Figure 4.8.



(a) frame S′ (b) frame S

Figure 4.8 The effect of relativistic beaming: (a) emission that is isotropic in the jet frame, S′, is seen in (b) the observer's frame, S, to be beamed into a narrow cone.

Relativistic beaming means that if an observer is located in the direction of this narrow cone of photons then the observed brightness will be very high, compared to what would be expected for a stationary object where the photons travel uniformly (isotropically) in all directions.

Since the angle of the cone of emission is small for speeds that approach c, $\tan\theta \sim \theta$ (in units of radians), and the expression for the angle into which the radiation with $\theta' < \pi/2$ is beamed simplifies to

$$\theta \approx \frac{1}{\gamma(V/c)} \approx \frac{1}{\gamma} \tag{4.8}$$

where the final approximation assumes $V/c \sim 1$. If you revisit your answers for Exercise 4.2 you should find that these approximations hold true for cases (ii) and (iii) where the angles are small and V approaches c.

4.1.3 Further boosting effects

As well as the beaming effect explained in the previous section, there are two other, closely related effects that can enhance the observed brightness of jets depending on their orientation: time dilation and spectral boosting.

Time dilation and photon arrival rate

The brightness of radiation that we measure at the Earth is related to the number of photons arriving in a given time interval, so luminosity is defined as the energy carried by the photons per unit time, while flux is the energy per unit time received per unit area. But the time interval we measure at the Earth between the arrival of photons is affected *both* by the 'raindrop' effect (the fact that we are effectively travelling towards the photons, which is equivalent to the traditional Doppler effect), and by time dilation.

The net result is that the time interval between the arrival time at Earth of two photons emitted with a separation of $\Delta t'_{\rm em}$ in the jet frame is given by

$$\Delta t_{\rm rec} = \frac{\Delta t'_{\rm em}}{\mathcal{D}} \tag{4.9}$$

where $\mathcal{D}$ is known as the **relativistic Doppler factor**, and is defined as

$$\mathcal{D} = \frac{1}{\gamma[1 - (V/c)\cos\theta_{\rm jet}]} \tag{4.10}$$

where $\theta_{\rm jet}$ is the angle between the jet's direction of travel and the line of sight (see Figure 4.5). Shorter time intervals between photons in the observer's frame means measuring a higher flux.

Spectral boosting

The relativistic Doppler factor relates the observed and emitted frequencies according to

$$\nu = \mathcal{D}\nu' \tag{4.11}$$

which is analogous to the non-relativistic Doppler effect. The reason that the frequency is affected is that, if we switch to thinking about the emitted radiation in its wave description, the argument above about time intervals also applies to the intervals between wavefronts *within* a packet of radiated energy, i.e. the photons we considered previously.

The shifting of frequency can also increase (or occasionally decrease) the brightness we observe, because it has the effect of changing which intrinsic frequency (wavelength) of radiation is being measured. You will see later that the radio emission from jets often has a decreasing power-law relationship between flux density and frequency so that $F_\nu \propto \nu^{-\alpha}$, where α, representing the slope of a log–log graph of the two quantities, is known as the **spectral index**.

These two ways in which time dilation enhances the brightness of a moving source relative to a stationary one are illustrated in Figure 4.9.

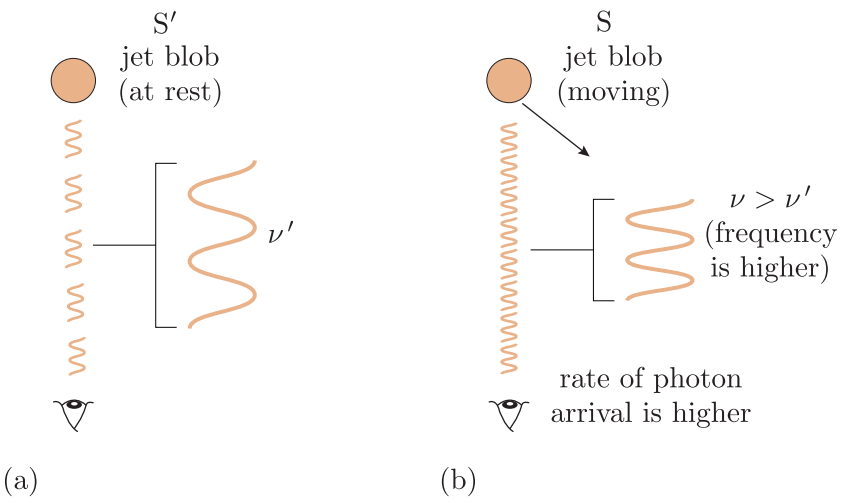


Figure 4.9 The effects of relativistic Doppler boosting: in the observer's frame S, both the flux of radiation (the rate at which photons arrive) and the frequency of the radiation are increased.

Total of all three relativistic boosting effects

Taking together all three of the effects we have discussed (beaming, time dilation and spectral boosting) the observed luminosity density (luminosity per unit frequency), L_ν, from a region of jet is related to the true, intrinsic luminosity density emitted by the jet, L'_ν, by

$$L_\nu = \mathcal{D}^{3+\alpha} L'_\nu \tag{4.12}$$

Exercise 4.3

The radio jet of the quasar 3C 273 is thought to be oriented with an angle of 6° to the line of sight. A bright jet knot is measured to have a luminosity density at a particular frequency of $3.7 \times 10^{24}\,\mathrm{W\,Hz^{-1}}$ and $\alpha = 0.6$, and the jet material is thought to be travelling at $V = 0.85c$. What is the intrinsic luminosity density emitted by the jet in its own frame of reference?

These relativistic beaming effects are very well studied, and the observed sidedness of radio jets is used to infer both speeds and orientation of the jets (i.e. θ_{jet}), to be able to model how the jets evolve and transfer energy to their surroundings.

It is important to note that all of these effects occur in other types of relativistic outflows as well. For example, relativistic jets are observed in X-ray binary star systems in which a stellar-mass black hole is powered by accreting material from a neighbour. In Chapter 5 of this book you will also see how these effects are important in gamma-ray bursts, which are powerful transient explosions linked to the endpoints of stellar evolution.

4.2 Matter and radiation in relativistic outflows

The previous section considered the large-scale motion of matter as it travels outwards in black-hole jets. As well as affecting the **macrophysics** of jets (processes affecting directly observable size scales), special relativity is *also* important for understanding the **microphysics** of how the radio emission we observe from jets is produced, e.g. how individual particles are behaving.

It was realised in the mid 1950s that both optical and radio jet emission must be produced by the mechanism of **synchrotron radiation**, which occurs when charged particles spiral around magnetic field lines at relativistic speeds. Figure 4.10a illustrates the basic process that produces synchrotron radiation, while Figure 4.10b places the microphysical radiation process in the context of the large-scale jet flow.

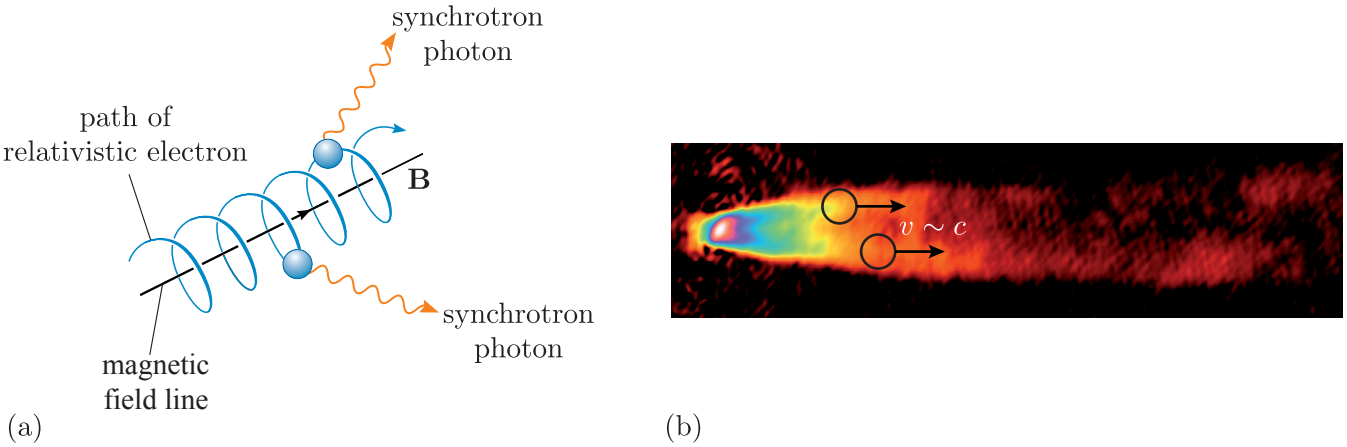


Figure 4.10 Two relativistic processes in jets: (a) the spiralling in a magnetic field **B** of individual charged particles at speeds close to c produces synchrotron radiation, and (b) the bulk relativistic flow of parcels of material (containing spiralling particles) along a jet.

In the next section we will explore the synchrotron process in more detail. As you work through the rest of the module it is very important to keep in mind the distinctions between jet macrophysics and microphysics. In particular, because both processes involve relativistic speeds (and corresponding Lorentz factors), it can be easy to get confused between the speed of the large-scale bulk motion of the jet fluid and the speeds of the individually spiralling electrons. In discussions of relativistic outflows we will always use γ for the Lorentz factor corresponding to the bulk velocity of jet material, and γ_e to refer to the typical Lorentz factors of individual spiralling particles (you will see shortly that the main particles of interest are electrons and positrons).

4.2.1 Synchrotron radiation

All of the electromagnetic radiation we measure is produced by the acceleration of charged particles in some form. In Chapter 3 you encountered thermal bremsstrahlung, which is caused by interactions between two charged particles. In the synchrotron process it is the interaction between particles and a magnetic field that accelerates particles and causes them to emit radiation.

A charged particle in a magnetic field will follow a spiral path around the magnetic field lines. The frequency of rotation is known as the **gyrofrequency**, which for a non-relativistic particle of mass m is given by

$$\nu_{\mathrm{g}} = \frac{|q|B}{2\pi m} \tag{4.13}$$

where B is the magnetic field strength and q is the particle's charge.

For a relativistic particle, the frequency of the synchrotron radiation produced by the spiral motion is related to ν_{g}. The emitted radiation from a single spiralling particle peaks at a frequency ν_{syn}, which is given by

$$\nu_{\mathrm{syn}} \approx \gamma_{\mathrm{e}}^2 \nu_{\mathrm{g}} = \frac{\gamma_{\mathrm{e}}^2 |q| B}{2\pi m} \tag{4.14}$$

Although Equation 4.14 is entirely general and applies to any charged particle, there is a reason for the inclusion of subscript 'e', which you will see shortly.

where γ_{e} is the Lorentz factor of the particle. In the case of non-relativistic particle speeds ($\gamma = 1$) the emission is known as cyclotron radiation. Figure 4.11 shows the synchrotron spectrum for an individual electron.

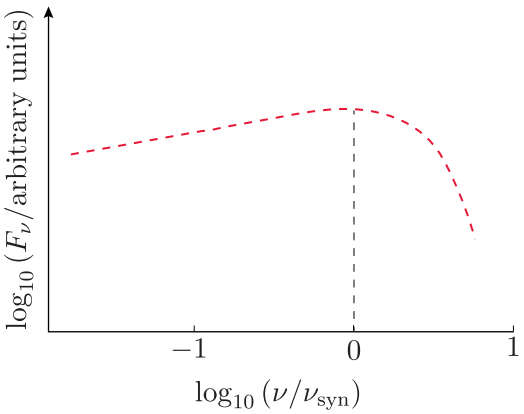


Figure 4.11 A log–log graph of flux density versus frequency for synchrotron radiation from a single electron.

The Lorentz factor of an individual relativistic particle is closely related to the particle's energy, as described in the following box.

Relativistic particle energies

The total energy E of a particle of mass m travelling at relativistic speeds is given by:

$$E = \gamma_{\mathrm{e}} m c^2 \tag{4.15}$$

The frequency at which synchrotron emission is produced depends strongly on the type of particle. Exercise 4.4 enables you to investigate how synchrotron radiation differs for electrons and protons.

Exercise 4.4

Calculate the approximate peak frequency of synchrotron radiation produced by (a) an electron and a proton, each with $\gamma = 1000$, and (b) an electron and a proton, each with a particle energy $E = 500\,\mathrm{MeV}$. In both cases assume a magnetic field strength of $B = 10^{-7}\,\mathrm{T}$.

The exercise shows that for similar Lorentz factors, electron synchrotron emission is produced at frequencies nearly 2000 times higher than those coming from a proton (reflecting the ratio of the particle masses). For similar particle *energies*, the difference is even more extreme. For typical magnetic field strengths in astrophysical situations, including in AGN jets, electron synchrotron emission is produced at GHz radio frequencies, whereas any proton contribution peaks at very low frequencies, below the range observable by radio telescopes.

The *rate* of synchrotron emission also depends strongly on γ_e and therefore inversely on particle mass. This means that the contribution of synchrotron emission from protons is always negligible compared to electrons. (Positrons, if any are present, will radiate at the same rate as electrons.) For the remainder of the book we will therefore consider only electron synchrotron radiation.

The amount of synchrotron radiation produced per unit time is the **synchrotron emissivity**, j_{syn}, which for a single relativistic electron of energy $E = \gamma_{\mathrm{e}} mc^2$ is given by its energy loss rate, as follows

$$j_{\mathrm{syn}}(E) = -\left\langle \frac{\mathrm{d}E}{\mathrm{d}t} \right\rangle = \frac{4}{3}\sigma_{\mathrm{T}}\gamma_{\mathrm{e}}^2 c \frac{B^2}{2\mu_0} \tag{4.16}$$

where σ_{T} is the Thomson cross-section and μ_0 is a constant known as the permittivity of free space. The rate of energy loss is given in angle brackets as it is an average loss rate on timescales longer than the gyration of the electron about the field lines. The synchrotron emissivity determines the flux we measure from a synchrotron source at the Earth, and so measurements of radio spectra from synchrotron sources can be used to test the predictions of synchrotron theory.

In reality we are never observing a single radiating electron, and so we need to consider the spectrum of a population of electrons, which are likely to have a range of energies. A typical electron energy distribution has the form of a power law, so that the number of electrons with energies in the range E to $E + \mathrm{d}E$ is given by:

$$N(E)\,\mathrm{d}E = N_0 E^{-p}\,\mathrm{d}E \tag{4.17}$$

where N_0 is a constant and p is known as the **electron energy index**.

The total emissivity in a given frequency range is then given by summing the spectra of all individual electrons that contribute within that range. A precise derivation of the synchrotron spectrum involves the integral of $j_{\rm syn} \times N(E)\,{\rm d}E$ multiplied by a convolution term that describes the shape of the individual electron's spectrum (i.e. the distribution in frequency shown in Figure 4.11). However, the correct shape of the synchrotron radio spectrum across much of the typically observed frequency range can be obtained by making the simplifying assumption that an electron of energy E emits all of its synchrotron emission at the relevant peak frequency, $\nu_{\rm syn}$.

The brightness quantity we measure with radio telescopes on Earth is the **flux density**, F_ν, measured in units of $\mathrm{W\,m^{-2}\,Hz^{-1}}$. This is equivalent to the flux measured in a very narrow frequency range, divided by that frequency range ${\rm d}\nu$. Example 4.4 considers the shape of this measured synchrotron spectrum.

Example 4.4

Consider a population of electrons that have a power-law distribution of particle energies, as given by Equation 4.17. Assuming that the observed flux in a given frequency range is proportional to the total emitted radiation over that range of frequencies, find an expression for how F_ν depends on ν and the magnetic field strength, B. (*Hint*: assume the source is at a redshift close to zero, so there is no difference between the observed and emitted frequencies.)

Solution

The flux we measure over a narrow frequency range ν to $\nu + {\rm d}\nu$ is given by $F_\nu\,{\rm d}\nu$ and is proportional to the product of the emissivity of electrons emitting in that range and the number of such electrons:

$$F_\nu\,{\rm d}\nu \propto j_{\rm syn}(E)\,N(E)\,{\rm d}E$$

where E is the energy of an electron whose emission peaks at frequency ν. We will work with proportionalities because the question asks only for the dependence on ν and B – the constant of proportionality would account for the distance to the object, i.e. the geometric conversion between emitted luminosity and flux.

Substituting in our expressions for $j_{\rm syn}$ and $N(E)\,{\rm d}E$ we obtain

$$F_\nu\,{\rm d}\nu \propto \frac{4}{3}\sigma_{\rm T}\gamma_{\rm e}^2 c\frac{B^2}{2\mu_0}N_0 E^{-p}\,{\rm d}E$$

Since we are working with a proportionality we can neglect the constant terms (except for B, which is constant, but the question asked us to retain; it will be relevant later in the chapter), and so if we substitute for $\gamma_{\rm e}$ in terms of E (Equation 4.15) we obtain

$$\begin{aligned} F_\nu\,{\rm d}\nu &\propto E^2 B^2 E^{-p}\,{\rm d}E \\ &\propto E^{2-p}B^2\,{\rm d}E \end{aligned}$$

We now need to rewrite this expression as a dependence on emitting frequency instead of particle energy, so we need expressions for how E and $\mathrm{d}E$ depend on ν and B. We first substitute for $\gamma_\mathrm{e} = E/mc^2$ in Equation 4.14 and then rearrange it to obtain

$$E = m_\mathrm{e}c^2 \left(\frac{2\pi m_\mathrm{e}\nu}{eB}\right)^{1/2}$$

and so $E \propto \nu^{1/2}B^{-1/2}$.

We can now differentiate the preceding expression to obtain a relation between $\mathrm{d}E$ and $\mathrm{d}\nu$:

$$\mathrm{d}E = m_\mathrm{e}c^2 \left(\frac{2\pi m_\mathrm{e}}{eB}\right)^{1/2} \frac{1}{2}\nu^{-1/2}\,\mathrm{d}\nu$$

so $\mathrm{d}E \propto \nu^{-1/2}B^{-1/2}\,\mathrm{d}\nu$. Substituting in the proportionalities for E and $\mathrm{d}E$ into the expression for $F(\nu)\,\mathrm{d}\nu$ gives

$$\begin{aligned} F_\nu\,\mathrm{d}\nu &\propto \left(\nu^{1/2}B^{-1/2}\right)^{2-p} B^2\nu^{-1/2}B^{-1/2}\,\mathrm{d}\nu \\ &\propto \nu^{-(p-1)/2}B^{(p+1)/2}\,\mathrm{d}\nu \end{aligned}$$

Finally, assuming a narrow frequency range we can divide by $\mathrm{d}\nu$ to obtain the required expression for the flux density, F_ν

$$F_\nu \propto \nu^{-(p-1)/2}B^{(p+1)/2} \tag{4.18}$$

Therefore if the underlying electron population has a power-law spectrum, then so will the observed synchrotron spectrum. Figure 4.12 illustrates how the spectra of individual electrons add up to produce an overall power-law shape across a wide range of frequencies.

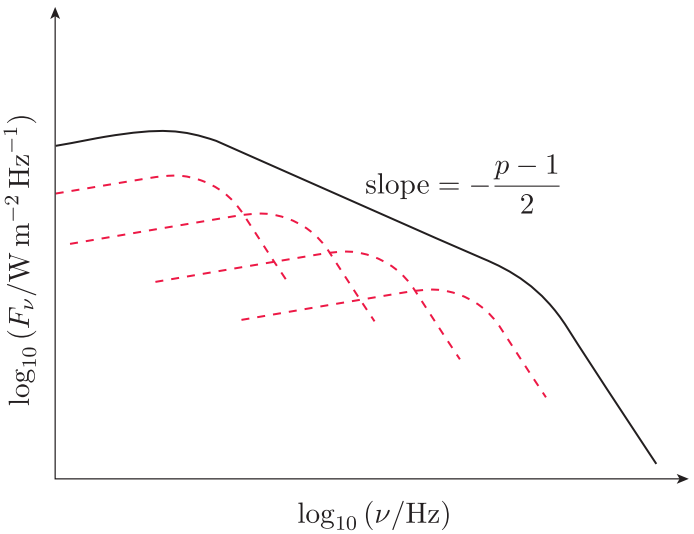


Figure 4.12 A sketch of the spectrum of synchrotron radiation from a population of electrons, obtained by summing the contributions from the individual electron spectra.

The spectral index, α, is defined as the observed spectral slope, and so it is related to the electron energy index via $\alpha = (p-1)/2$. In the next section we will explore why the underlying electron energy distributions typically have the power-law form given by Equation 4.17.

4.2.2 Particle acceleration and shocks

The conclusion that radio emission from radio galaxies and radio-loud quasars is produced by the synchrotron process tells us something important about the physics of jets: they must contain very energetic particles. In this section we will consider how these energetic particles are produced.

The basic idea is that particles are accelerated at **shocks**, which are discontinuities in the fluid flow caused by a disturbance attempting to travel at speeds faster than the local sound speed.

If we consider fluid flow in a jet, it is very likely that this is not entirely smooth – parcels of plasma are likely to be ejected from the central regions at varying rates, and so catch-up and collision are possible. Figure 4.13a illustrates a pressure wave (a discontinuity in the gas) travelling through the jet plasma. If its speed exceeds c_s then it becomes a shock. In this situation, the unshocked gas has no warning of the peak of the disturbance arriving, which leads to an abrupt discontinuity in the fluid properties on either side of the shock front, as shown in Figure 4.13b.

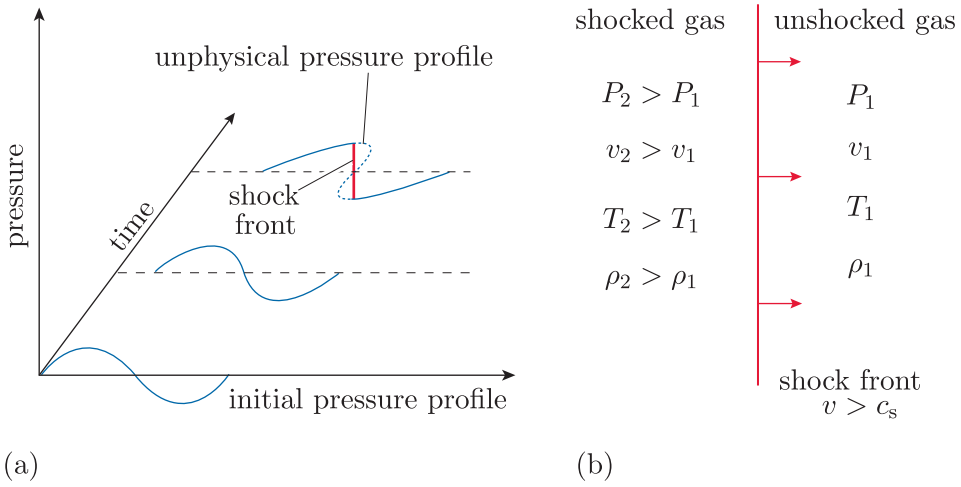


Figure 4.13 (a) A propagating pressure disturbance that turns into a shock because the wavefront gets compressed to form an abrupt jump in pressure; (b) gas properties differing abruptly on either side of the shock front.

Recalling our discussion of microphysical versus macrophysical processes, if we now shift back to considering the behaviour of individual particles instead of larger parcels of gas, it turns out that individual electrons can travel backwards and forwards across shock fronts because they continue to follow their spiralling paths around magnetic field lines.

When a particle crosses the shock front then in the particle's rest frame it sees the plasma moving towards it, and typically gains energy by scattering off magnetic field concentrations in (what it sees as) the approaching flow. Crucially, this is true whichever direction the particle travels across the front, and so particles can scatter backwards and forwards, with an average energy increase ΔE per crossing given by

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4v}{3c} \tag{4.19}$$

where v is the difference in the speed of the fluid on either side of the shock front. From this equation it can be seen that this process only works efficiently for high-energy particles (i.e. those whose speeds are already relativistic).

This process is known as **diffusive shock acceleration** or **(first-order) Fermi acceleration**, and occurs not just in jets but in a wide range of astrophysical environments. The following study comment shows how the particle population can evolve to a power-law distribution of energies.

The energy distribution of shock-accelerated particles

Consider a population of $N_{\rm init}$ particles in the vicinity of a shock front that all have the same energy, $E_{\rm init}$. We assume that a particle has a non-zero probability, q, of remaining near the shock front after each round trip across the shock front and back. We write the energy increase per round trip as $\epsilon = 1 + \Delta E/E$, where $\Delta E/E$ is given by Equation 4.19.

The number of particles remaining after one round trip will be $N_1 = N_{\rm init}q$. For each successive round trip, the particle number is again multiplied by q, so that after n round trips there will be $N_{\rm n} = N_{\rm init}q^n$ particles remaining. The energy of each particle after one round trip will be $E_1 = E_{\rm init}\epsilon$, and so similarly, after n round trips, the particles will typically have an energy $E_{\rm n} = E_{\rm init}\epsilon^n$.

We can combine the equations for $N_{\rm n}$ and $E_{\rm n}$ to eliminate n. Taking the natural log of the equations for $N_{\rm n}$ and $E_{\rm n}$ gives

$$n = \frac{\ln(N_{\rm n}/N_{\rm init})}{\ln q} = \frac{\ln(E_{\rm n}/E_{\rm init})}{\ln \epsilon}$$

Rearranging, this gives

$$\frac{\ln\left(N_{\rm n}/N_{\rm init}\right)}{\ln\left(E_{\rm n}/E_{\rm init}\right)} = \frac{\ln q}{\ln \epsilon}$$

which can be rearranged further to give

$$\frac{N_{\rm n}}{N_{\rm init}} = \left(\frac{E_{\rm n}}{E_{\rm init}}\right)^{\ln q/\ln \epsilon} \tag{4.20}$$

So far we are considering only the particles that remain circling the shock front after a certain number of round trips, but when we observe a source we are interested in the energy distribution of all of the particles, including those that escape at different times.

At a given point in time, N_n describes the number of particles that will eventually end up with energies *greater* than E_n, because some will leave before gaining any more energy, while some will continue to reach higher energies. So N_n can be written as

$$N_n = \int_{E_{init}}^{\infty} N(E)\,dE \propto E^{\ln q/\ln \epsilon}$$

where $N(E)\,dE$ is the number of particles with energies between E and dE. (We use this expression rather than simply $N(E)$ because it is more realistic to consider numbers within a narrow energy range, dE, rather than a single fixed energy E.)

Evaluating the integral gives

$$N(E)\,dE \propto E^{-p}\,dE$$

where $p = -1 + \ln q/\ln \epsilon$, which is equivalent to the electron energy index defined in Equation 4.17. It can be shown that for the expected conditions at a strong shock, $p \sim 2$.

- If the particles in radio jets have been shock-accelerated by the process described in the preceding box, what would you expect to measure as the typical radio spectral index, α?

- If $p = 2$ then $\alpha = (p-1)/2 = 0.5$.

Observations show that many jets and hotspot regions of radio galaxies, where we expect particles to be accelerated at shocks, indeed show $\alpha \approx 0.5$. The α values for more extended regions of radio emission, such as the largest-scale plumes and lobes in Figures 4.1 and 4.2, tend to be steeper as a result of more complicated effects of how particle energies evolve over long timescales once acceleration stops.

4.3 Energetics and galaxy feedback

In the final section of the chapter we return to broader questions of how jets are produced, and how they transport energy within and beyond their host galaxies, as introduced at the end of the previous chapter.

4.3.1 Powering AGN jets

In order for jets of matter to travel at relativistic speeds for distances of hundreds of kiloparsecs, a large amount of energy must be channelled into a narrow outflowing region. This energy supply must remain quite stable on a timescale of tens of millions of years.

The energy supply and launching of narrow, highly relativistic jets is tightly connected to the supermassive black holes in galaxy centres. The favoured mechanism by which jets are launched is the **Blandford–Znajek mechanism**, in which a spinning black hole twists magnetic field lines that produce a channel through which electromagnetic energy is transported outwards from close to the black hole. Figure 4.14 shows a diagram of how this mechanism works.

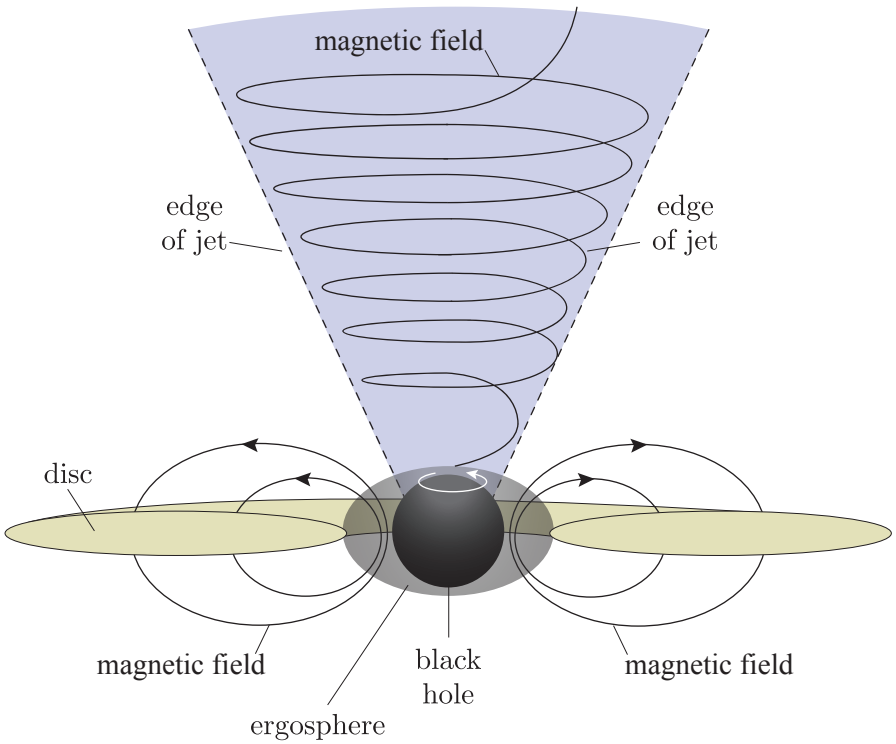


Figure 4.14 A schematic illustration of the Blandford–Znajek mechanism for launching relativistic jets. The diagram shows a side on view of an accretion disc and spinning black hole, with jets launched in the direction of the black-hole spin axis (assumed perpendicular to the disc of accreting material).

A crucial requirement is that the black hole is spinning. A spinning black hole has an **ergosphere**, a region outside the event horizon where spacetime is forced to rotate; it is this region that allows the twisted magnetic field structure to be produced. If the Blandford–Znajek mechanism is the correct explanation for how jets are produced, then the spin of the central black hole might determine which AGN become radio-loud and which do not.

The second requirement for jets to be launched is that energy is supplied via accretion, which transports magnetic field inwards to the ergosphere. A rough estimate of the rate at which energy released by accretion is available to power a jet can be written as

$$Q_{\rm jet} = \eta_{\rm jet}\dot{m}c^2 \tag{4.21}$$

where $\dot{m}$ is the rate at which mass is accreted and $\eta_{\rm jet}$ is the efficiency with which mass is converted to energy. An efficiency of 1 would correspond to all of the mass–energy of the accreting material being converted to available energy.

A similar expression was given in Chapter 1 for the rate of energy release via radiation, i.e. the AGN luminosity:

$$L_{\rm AGN} = \eta_{\rm rad}\dot{m}c^2 \tag{4.22}$$

where the efficiency $\eta_{\rm rad}$ is likely to be different from $\eta_{\rm jet}$, and they cannot together sum to more than 1. Typically $\eta_{\rm rad}$ is thought to be $\ll 1$, e.g. $\eta_{\rm rad}$ is usually assumed to be ~ 0.1 for bright AGN (although it can be much lower in some situations).

The following exercise explores the relationship between accretion and the energy budget of an active galaxy with powerful jets.

Exercise 4.5

An AGN is observed to have jets of power $Q_{\rm jet} = 3.5 \times 10^{38}$ W as well as having a bright AGN nucleus. If 40% of the energy released from accretion onto the central black hole goes into powering the jet, and assuming a typical radiative efficiency of $\eta_{\rm rad} = 0.1$, estimate the accretion rate of the black hole in units of $\mathrm{M_\odot\,y^{-1}}$, and the luminosity of the AGN.

4.3.2 Energy content of radio galaxies

When radio jets were first detected, attempts were made to estimate how much energy they contain as a first step in understanding how they could be produced – the required energies turned out to be very large, which is why jets are thought to have important effects on the evolution of their galaxy environments.

The total energy density (energy per unit volume, $U_{\rm tot}$) contained within a region of synchrotron-emitting plasma, such as the lobes of a radio galaxy, is the sum of the energy densities of the particles and the magnetic field:

$$U_{\rm tot} = \int_{E_{\rm min}}^{E_{\rm max}} En(E)\,{\rm d}E + \frac{B^2}{2\mu_0} + U_{\rm NR} \tag{4.23}$$

where the integral term describes the energy from synchrotron-radiating electrons. We call this integral $U_{\rm e}$ from now on. The function $n(E)$ is the number density of electrons of energy E, so is given by $N(E)$ (as defined in

Equation 4.17) divided by the source volume, V. The second term is the magnetic field energy density. $U_{\rm NR}$ is the energetic contribution from any 'non-radiating' particles that are present but not contributing to the synchrotron emission, e.g. protons – see Section 4.2.1. The integral limits $E_{\rm min}$ and $E_{\rm max}$ are the minimum and maximum energies of synchrotron-emitting electrons. The total energy of the radio galaxy is then $E_{\rm tot} = U_{\rm tot}V$.

Exercise 4.6

Find an expression for the electron energy density, $U_{\rm e}$, in the situation where $n(E) = n_0 E^{-p}$ and (i) $p = 2$; and (ii) $p \neq 2$.

Although the equation for the total energy density is relatively straightforward, it is difficult in practice to measure this quantity reliably. Firstly, an assumption must be made about $U_{\rm NR}$, because it cannot be determined from observations (since if protons are present we can't detect their radiation). The ratio of non-radiating to radiating particles is typically represented by κ, in which case Equation 4.23 can be written as

$$U_{\rm tot} = (1+\kappa)\int_{E_{\rm min}}^{E_{\rm max}} En(E)\,{\rm d}E + \frac{B^2}{2\mu_0} \tag{4.24}$$

The second difficulty in measuring $U_{\rm tot}$ is that the quantity we can measure, the synchrotron flux density, depends on both the electron energy distribution *and* the magnetic field strength. Equation 4.18 can be rewritten as

$$F_\nu(\nu) = Cn_0\nu^{-(p-1)/2}B^{(p+1)/2} \tag{4.25}$$

where C is a constant and the dependence on n_0 (also a constant) is written out explicitly. This equation tells us that, while the radio spectral index can be used to determine p, the measured radio flux density at a particular frequency only tells us a combination of n_0 and B, rather than each quantity separately: a small electron density and high magnetic field can produce the same amount of emission as a high electron density and small magnetic field.

However, the following example shows that it is possible to determine a *minimum* total energy contained within the plasma, which provides useful information about the jet power and how it could be generated.

Example 4.5

Show that, for a measured flux density F_ν at a frequency ν, there is a minimum total energy that the plasma could contain, and show that the magnetic field strength at which this occurs, $B_{\rm min} \propto F_\nu^{2/(p+5)}$.

(*Hint*: to simplify the algebra, consider the situation where $p \neq 2$.)

Solution

The strategy to find a minimum energy must involve finding the conditions for which the derivative of the total energy with respect to some quantity is equal to zero. The quantity in question must be B, since we are asked to find the value of B corresponding to the minimum.

The first step is to write the expression for the total energy in terms of quantities related to F_ν. Expanding out Equation 4.24, as was done in Exercise 4.6, gives

$$U_{\rm tot} = (1+\kappa)\frac{n_0}{(2-p)}\left(E_{\rm max}^{2-p} - E_{\rm min}^{2-p}\right) + \frac{B^2}{2\mu_0}$$

We are aiming for an expression that depends on F_ν and B, and so we can rearrange Equation 4.25 for n_0.

$$n_0 = \frac{F_\nu}{C}\nu^{(p-1)/2}B^{-(p+1)/2}$$

We can now use this expression to eliminate the unknown n_0 from the equation for $U_{\rm tot}$:

$$U_{\rm tot} = (1+\kappa)\frac{F_\nu \nu^{(p-1)/2}B^{-(p+1)/2}}{C(2-p)}\left(E_{\rm max}^{2-p} - E_{\rm min}^{2-p}\right) + \frac{B^2}{2\mu_0}$$

Although this is a complicated expression, it is relatively straightforward to differentiate it with respect to B:

$$\frac{{\rm d}U_{\rm tot}}{{\rm d}B} = (1+\kappa)\frac{F_\nu \nu^{(p-1)/2}}{C(2-p)}\left(E_{\rm max}^{2-p} - E_{\rm min}^{2-p}\right)\frac{-(p+1)}{2}B^{-(p+3)/2} + \frac{B}{\mu_0}$$

Setting the derivative to zero and rearranging for B gives

$$B_{\rm min} = \left[\mu_0(1+\kappa)\frac{F_\nu \nu^{(p-1)/2}(p+1)}{C(4-2p)}\left(E_{\rm max}^{2-p} - E_{\rm min}^{2-p}\right)\right]^{2/(p+5)} \tag{4.26}$$

and so the magnetic field strength is proportional to $F_\nu^{2/(p+5)}$, as required.

It can be shown that this **minimum energy condition** for a synchrotron plasma corresponds very closely to the condition of having equal energy densities in the magnetic field and in the particles, which is known as **equipartition**, and given by

$$\frac{B^2}{2\mu_0} = (1+\kappa)\int_{E_{\rm min}}^{E_{\rm max}} En(E)\,{\rm d}E \tag{4.27}$$

where κ is the ratio of non-radiating to radiating particles.

The ability to measure the minimum total energy in radio galaxies is really useful, because it tells us that the jet power must be sufficient to provide at least that much energy over the lifetime of the radio source. The jet must also have had the power to push the surrounding medium out of the way to expand the radio lobe. A useful quantity is therefore the **enthalpy**

of the radio galaxy, H, which is the sum of the internal energy and the work done to expand to its current size: $H \approx E_{\rm tot} + P_{\rm ext}V$.

This definition uses the simplifying assumptions that the pressure has remained the same throughout the expansion of the radio plasma, and that the lobe expansion has mainly been subsonic. It is likely that in some cases much of the expansion was supersonic, in which the enthalpy is higher, and so the energy that must have been supplied to an observed radio galaxy over its lifetime, $E_{\rm RG}(= H)$, is

$$E_{\rm RG} \geq E_{\rm tot} + P_{\rm ext}V \tag{4.28}$$

The following exercise explores a practical example.

Exercise 4.7

A radio galaxy has two lobes, with the overall distribution of radio plasma supplied by the two jets to the lobes having a roughly cylindrical shape. The overall length of the source from end to end is 500 kpc, and the lobes are roughly 30 kpc in radius. The radio galaxy is located in a galaxy cluster, with a typical ICM pressure of 1.7×10^{-13} Pa. Assume that the radio galaxy is at equipartition, and has a magnetic field strength of $B = 4 \times 10^{-9}$ T.

(a) Estimate the total energy contained within the radio lobes.

(b) Estimate the enthalpy of the radio galaxy ($E_{\rm RG}$) at the time of observation.

(c) If the radio galaxy has an age of 10^8 y, what jet power would have been required to produce the observed radio source (assuming $Q_{\rm jet}$ has been constant throughout its lifetime)?

The previous exercise shows how radio observations can be used to estimate the energy available from radio galaxies. These types of calculation helped establish the presence of a central supermassive black hole in radio galaxies and quasars, because other potential energy sources could not provide the necessary power.

As you saw in the previous chapter, understanding the energy being transported by radio galaxies is also very important for investigating how gas cools to form stars: jet energy transport has a strong influence on how massive galaxies evolve, and calculations such as those shown here enable galaxy feedback models to be tested.

4.4 Summary of Chapter 4

- Galaxies possessing large-scale, radio-emitting jets are known as **radio galaxies** and **radio-loud quasars**, which are part of the **active galaxy** population.
- Active-galaxy jets are produced in the vicinity of a central supermassive black hole, and travel at relativistic speeds. Such jets can extend for distances of over a megaparsec – well beyond the boundary of the host galaxy.
- If orientation on the sky is not taken into account then apparent jet speeds can appear to exceed the speed of light, a phenomenon known as **apparent superluminal motion**. The measured speed is related to the true jet speed by

$$\beta_{\rm app} = \frac{\beta \sin\theta}{1 - \beta\cos\theta} \qquad \text{(Eqn 4.3)}$$

- The relativistic jet speeds mean that measurements of jet properties are subject to effects of special relativity, including **relativistic beaming**, in which isotropically emitted radiation is observed in a narrow cone around the direction of motion, luminosity and spectral boosting due to the relativistic Doppler effect.
- At a given observing frequency, the luminosity density, L_ν, of a jet region is boosted relative to the emitted luminosity density at that frequency, L'_ν, according to

$$L_\nu = \mathcal{D}^{3+\alpha} L'_\nu \qquad \text{(Eqn 4.12)}$$

  where α is the radio spectral index and $\mathcal{D}$ is the **relativistic Doppler factor**, given by

$$\mathcal{D} = \frac{1}{\gamma[1 - (V/c)\cos\theta_{\rm jet}]} \qquad \text{(Eqn 4.10)}$$

  where γ is the Lorentz factor, and $\theta_{\rm jet}$ is the angle between the jet's direction of travel and the line of sight.
- The radio emission from the jets and lobes of radio galaxies (as well as many other astrophysical sources) is produced via the process of **synchrotron radiation**, in which relativistic electrons spiral around magnetic field lines.
- The synchrotron emission from a single relativistic electron peaks at a frequency $\nu_{\rm syn}$:

$$\nu_{\rm syn} \approx \frac{\gamma_{\rm e}^2 |q| B}{2\pi m} \qquad \text{(Eqn 4.14)}$$

  and the emissivity is given by

$$j_{\rm syn}(E) = \frac{4}{3}\sigma_{\rm T}\gamma_{\rm e}^2 c \frac{B^2}{2\mu_0} \qquad \text{(Eqn 4.16)}$$

- The electron populations in radio galaxy jets and lobes typically have a power-law distribution of energies

$$N(E)\,\mathrm{d}E = N_0\,E^{-p}\,\mathrm{d}E \qquad \text{(Eqn 4.17)}$$

  where the **electron energy index** p is related to the observed **spectral index**, α, by $\alpha = (p-1)/2$.

- The relativistic particles in radio galaxy jets and lobes have been accelerated at **shocks** caused by disturbances propagating at speeds greater than the local sound speed – this process results in the observed power-law energy distribution.

- Relativistic jets are thought to be produced by the **Blandford–Znajek mechanism** and powered by the accretion of gas onto a supermassive black hole. The rate at which a jet can be powered can be expressed as

$$Q_{\mathrm{jet}} = \eta_{\mathrm{jet}}\dot{m}c^2 \qquad \text{(Eqn 4.21)}$$

  where η_{jet} is the efficiency and $\dot{m}$ is the mass accretion rate.

- The total internal energy of a synchrotron-emitting plasma is given by the volume of the emitting region multiplied by the energy density, U_{tot}:

$$U_{\mathrm{tot}} = (1+\kappa)\int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} E\,n(E)\,\mathrm{d}E + \frac{B^2}{2\mu_0} \qquad \text{(Eqn 4.24)}$$

  where κ is the ratio of non-radiating to radiating particles and $n(E)$ is the electron number density, whose energy dependence is the same as for $N(E)$.

- The **minimum energy condition** or assumption of **equipartition** of energy densities between the particles and the magnetic field allows the internal energy of radio jets and lobes to be estimated.

- The total energy that the jets must have provided over a radio-galaxy's lifetime to explain its currently observed properties is given by

$$E_{\mathrm{RG}} \geq E_{\mathrm{tot}} + P_{\mathrm{ext}}V \qquad \text{(Eqn 4.28)}$$

Chapter 5 Gamma-ray bursts

In November 2004, a NASA Delta 7320 rocket blasted off from Cape Canaveral in Florida carrying a new space telescope called the *Swift Gamma-Ray Burst Explorer* (or just *Swift* for short) into low Earth orbit.

In 2018, *Swift* was renamed the *Neil Gehrels Swift Observatory* in honour of the scientist who led the project to build and launch the telescope.

Roughly three times per fortnight, the Burst Alert Telescope (BAT) on board the *Swift* satellite detects a bright flash of γ-rays arriving from space. These events are called γ-ray bursts (or GRBs for short). They produce a flash of γ-rays that only lasts a few seconds, but can be so bright that it briefly outshines all other γ-ray emission in the sky, including the Sun!

Our current understanding of GRBs comes from observations and diligent research spanning more than 50 years. For most of that time there was no certainty about the celestial origins of GRBs or the physical processes that produced them. However, the twenty years following the launch of *Swift* have seen huge advances in our knowledge of these spectacular phenomena.

In this chapter you will learn about the observed properties of GRBs and how these properties can be used to infer the astrophysical processes that produce them. You will see how observations spanning the entire electromagnetic spectrum are used to build physical models of GRBs as relativistic outflows, with some similar underlying physics to the jets studied in Chapter 4, but very different celestial origins.

Objectives

Working through this chapter will enable you to

- describe what is meant by the 'prompt emission' from GRBs and summarise its observational characteristics including properties of γ-ray light curves and spectra
- name the two different categories of GRB and explain how they are distinguished by their observational characteristics
- describe what is meant by the 'afterglow' of a GRB and summarise the typical behaviour of GRB afterglow light curves
- explain how the observed characteristics of GRBs can be used to infer that they must contain highly relativistic outflows
- describe the fireball model of GRB emission and explain how it accounts for the observed properties of GRBs' prompt and afterglow emission
- describe the different types of celestial event that are believed to be the progenitors of different GRB classes and outline the evidence for these associations.

5.1 GRB observations: a brief history

The history of GRB research began on 2 July 1967, at the height of the Cold War. The first GRB detection was made by two satellites that had been launched by the United States of America to monitor nuclear detonations in space around the Earth (or potentially behind the Moon) that would violate the Nuclear Test Ban Treaty. These satellites, named Vela 3 and Vela 4, detected a flash of γ-rays that did not resemble any of the expected signatures of a nuclear explosion. News of the detection did not reach the scientific community until 1973, but it was soon established that the fleet of Vela satellites had detected several of these flashes that seemed to appear at random times and to originate from random directions on the sky.

The next two decades saw relatively few new GRB observations, and the sparsity of observational data was a major obstacle to theoretical work to understand the origins of these mysterious flashes. This hiatus came to an end in 1991 with the launch of NASA's *Compton Gamma Ray Observatory* (*CGRO*) which carried a dedicated all-sky γ-ray telescope called the Burst and Transient Source Explorer (or BATSE for short). Over the next nine years the *CGRO* revolutionised the field of GRB research. BATSE detected an average of one GRB per day and was able to measure the bursts' light curves, their spectra, and their locations on the sky.

Even with the wealth of observational data that were collected by BATSE, several major questions about the nature and origins of GRBs remained outstanding. The next revolution in our understanding of GRBs came in 1997 when the first, rapidly fading, optical counterpart of a GRB was detected. It was quickly realised that studying these counterparts would be crucial to unveil the physical processes that operate in GRBs; you will learn more about them in Section 5.2.2.

Swift was specifically designed with instruments that could both detect GRBs and conduct rapid follow-up observations at X-ray, ultraviolet and optical frequencies. Like its predecessor, the *CGRO*, *Swift* dramatically improved our understanding of GRBs and it continues to operate and detect new bursts at the time of writing in 2023. You will learn more about the capabilities of *Swift* and the instruments it carries in Section 5.2.2.

In 2008, NASA launched the *Fermi* space telescope. *Fermi* carries two γ-ray telescopes called the Gamma-ray Burst Monitor (or GBM for short) and the Large Area Telescope (or LAT).

Figure 5.1 shows the celestial positions of all the GRBs detected by the *CGRO* BATSE, *Swift* and *Fermi* (up to 2023).

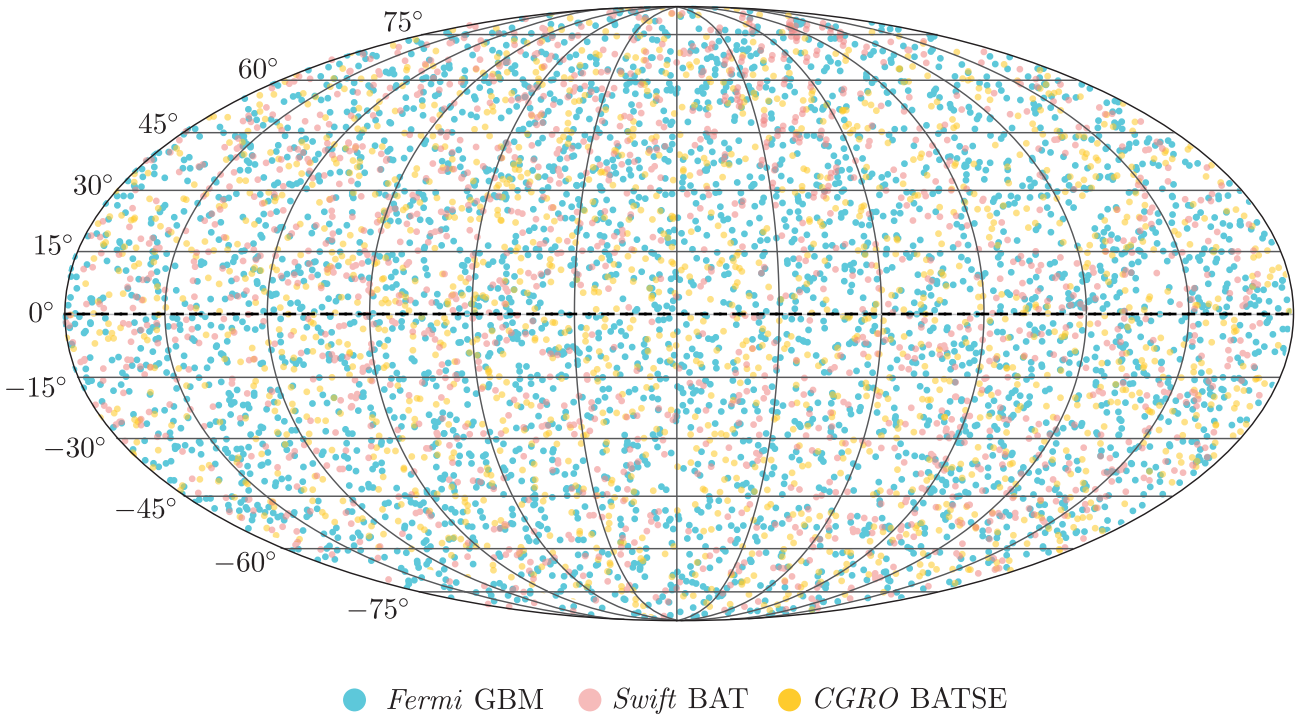


Figure 5.1 The spatial distribution of all the GRBs that were observed by the *CGRO* BATSE, *Fermi* GBM and *Swift* BAT instruments in Galactic celestial coordinates. The dashed line shows the location of the galactic plane.

Until recently, *Fermi* held the record for the highest energy γ-ray photon that had ever been detected coming from a GRB. This record stood at 94 million electronvolts (or about 1.5×10^{-11} J) until the ground-based, MAGIC γ-ray telescopes on the island of La Palma detected several γ-ray photons with energies above 1 billion electronvolts!* The fact that GRBs emit γ-rays with such enormous energies tells us that extreme physical processes must be involved.

We will end this section with an observational signature of a GRB that did not involve photons of *any* energy. In 2017, the LIGO and Virgo interferometers that were briefly introduced in *Cosmology* Chapter 3 detected gravitational waves produced in the seconds before the first γ-rays from the associated burst were emitted. In Section 5.5.2 you will learn how these and subsequent detections have allowed scientists to locate the progenitors of some GRBs and to finally unlock one of the longest-standing mysteries in astronomy.

*For context 1 billion electronvolts is roughly equal to the gravitational potential energy released when a grain of sand falls 1 millimetre at the Earth's surface. This may seem small, but remember this is the energy of a single photon and we can describe it in terms of objects and distances that we can experience physically every day!

5.2 Observable properties of GRBs

In the years since 1967, astronomers have observed and catalogued the properties of thousands of GRBs. In this section we will review those observational properties and you will see that GRBs exhibit a remarkable diversity of temporal and spectral behaviour.

5.2.1 The prompt-emission phase

At photon energies exceeding 10 keV, the bright γ-ray emission from GRBs is actually remarkably brief. Even the longest-lasting GRBs only remain detectable for a few tens of minutes. Astronomers call this transient burst of γ-rays the **prompt-emission phase**.

In the next three sections we will discuss the different temporal and spectral behaviours that are observed in GRBs and how they can be used to classify individual GRBs. We end our discussion of the prompt-emission phase with a discussion of the large-scale spatial distribution of GRBs.

Prompt GRB light curves

The observed durations of prompt GRB emission span a very large range, from a few milliseconds to tens of minutes. The detailed evolution of the γ-ray emission intensity during the prompt-emission phase (sometimes called the light curve *morphology*) also varies markedly between individual bursts. To illustrate this remarkable diversity of temporal behaviour, the panels in Figure 5.2 show prompt γ-ray light curves for eight GRBs that were detected by the *Swift* BAT instrument. The light curves show a wide range of durations and shapes, but all exhibit rapid variability on ms timescales. The gap in the light curve shown in panel (f) represents a short break in data collection by the BAT.

As these examples show, GRB light curves have complex shapes, with some having multiple bright emission episodes. These episodes may or may not be separated by gaps during which no γ-ray emission is detected. Some GRB light curves exhibit a single episode of continuous γ-ray emission which makes them appear much simpler. Even then, the rate at which the γ-ray flux increases and decays varies between bursts making some light curves appear symmetric, while others look more skewed. One temporal characteristic that does seem to be similar among all GRBs is the presence of very rapid variability, and flux variation on timescales as short as 1 millisecond has been detected from bright GRBs.

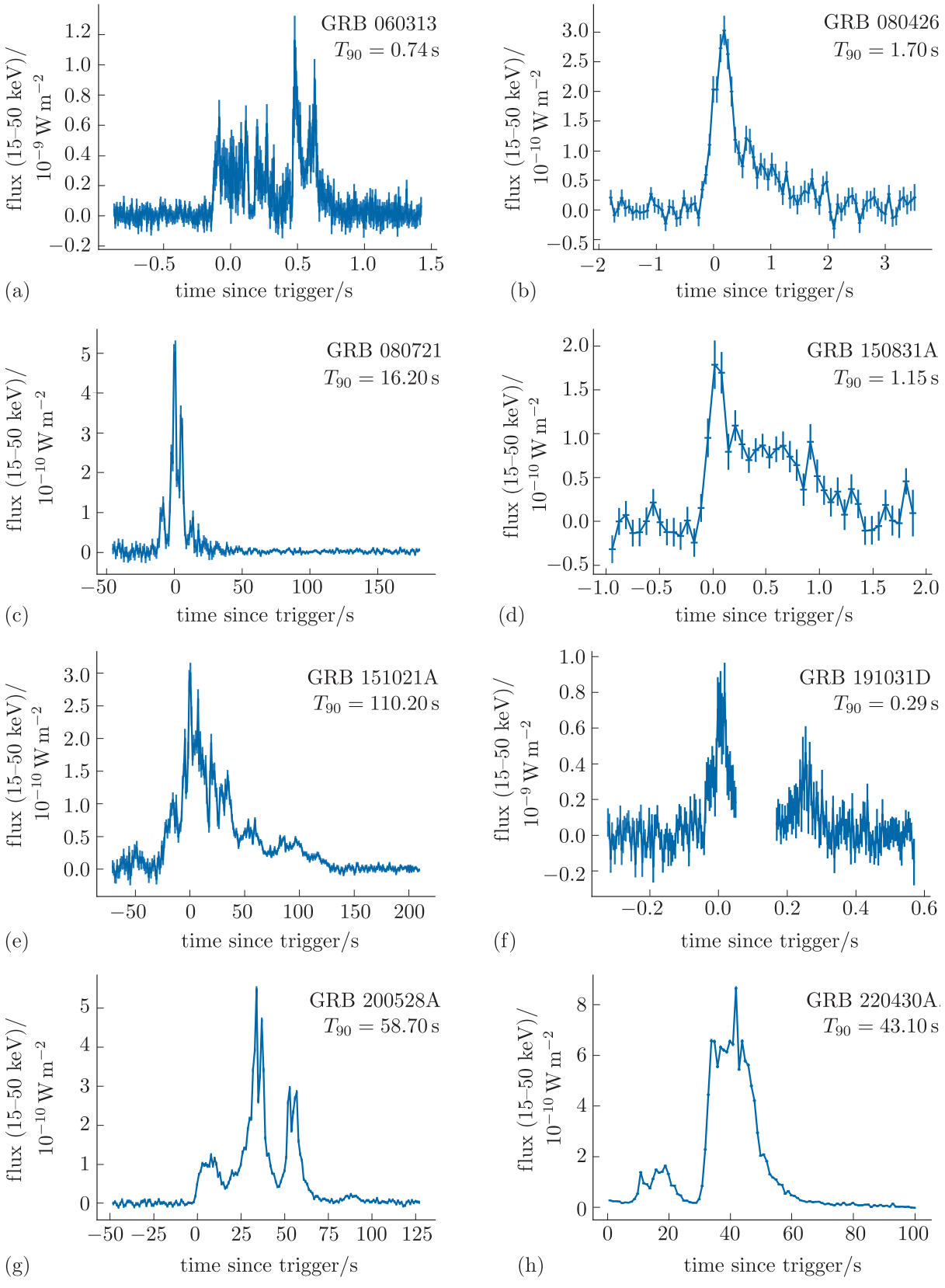


Figure 5.2 Eight prompt γ-ray light curves that were recorded by the *Swift* BAT instrument. See the box titled 'Comparing GRB durations' for an explanation of how the T_{90} value above each light curve is derived.

The rapid variability of GRBs means that it could be misleading to measure their brightness at any particular instant in time. Instead, the brightness of individual GRBs is normally specified in terms a time-integrated quantity called the fluence.

Fluence

The **fluence** S of a celestial object is defined as the observed flux F integrated over time

$$S = \int_{t_1}^{t_2} F(t)\,\mathrm{d}t \tag{5.1}$$

Sometimes you will see notation that defines the fluence for a range of frequencies. For example $S_{15\text{–}150\,\mathrm{keV}}$ denotes the observed fluence of photons with energies between 15 and 150 keV. Fluence has SI units of $\mathrm{J\,m^{-2}}$.

Observed GRB fluences vary widely from burst to burst. For example, the *Swift* BAT instrument measured values for $S_{15\text{–}150\,\mathrm{keV}}$ spanning four orders of magnitude between 10^{-11} and $10^{-7}\,\mathrm{J\,m^{-2}}$.

Comparing GRB durations

Instead of comparing the *total* durations of GRBs, astronomers will often measure the time interval during which a certain percentage of the total fluence was observed. The most commonly compared interval is denoted T_{90}, where the subscript 90 indicates that this is the duration within which 90% of the total fluence was observed. You may also see shorter intervals, like T_{50}, which represents the time interval over which half the fluence was observed.

Prompt GRB spectra

The prompt emission from GRBs spans a wide range of photon energies, with observed spectra extending from $\sim$10 keV to as much as $\sim$1 TeV in some extreme cases! As an example, Figure 5.3 shows the prompt spectral energy distribution for GRB 180720B, which is one of the brightest bursts detected by the *Fermi* LAT.

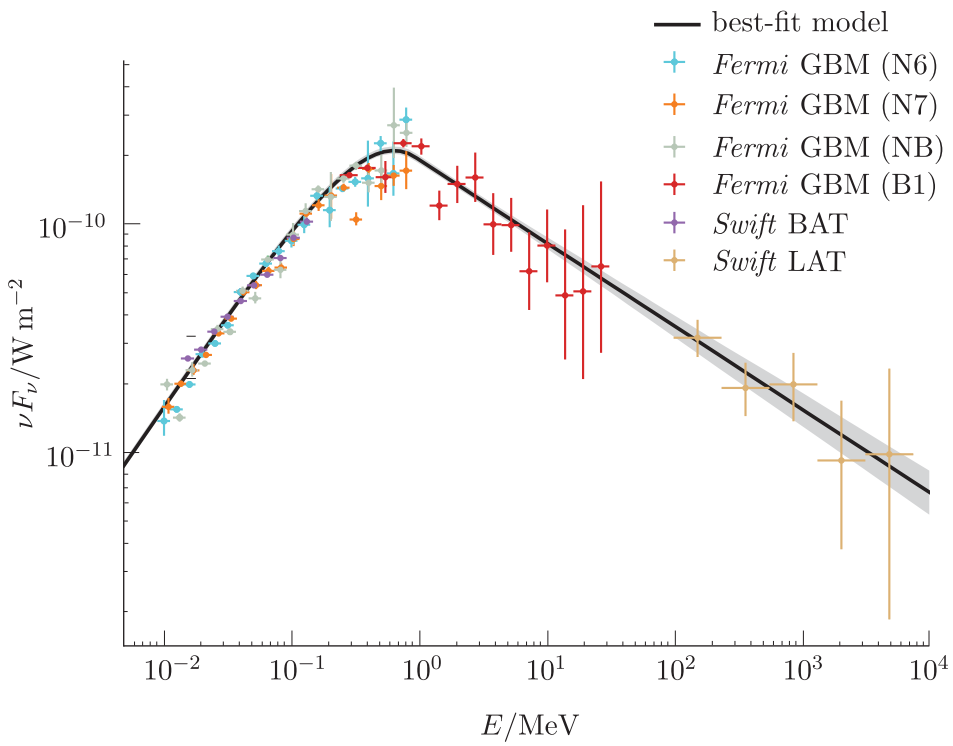


Figure 5.3 The prompt spectral energy distribution for GRB 180720B, which has $\nu_\mathrm{p} \approx 0.8\,\mathrm{MeV}$. The coloured points show the measurements made by instruments on board the *Swift* and *Fermi* satellites. The black line shows the best fit for a smoothly broken power-law model and the grey band gives its associated uncertainty.

Even though GRB 180720B was particularly bright at GeV energies compared to other GRBs, the overall *shape* of its spectrum is very typical. The prompt spectra of almost all GRBs can be roughly modelled using a broken power-law function

$$N_\nu \,\mathrm{d}\nu \propto \begin{cases} \left(\dfrac{\nu}{\nu_\mathrm{p}}\right)^{\alpha} \mathrm{d}\nu & \text{if } \nu \leq \nu_\mathrm{p} \\ \left(\dfrac{\nu}{\nu_\mathrm{p}}\right)^{\beta} \mathrm{d}\nu & \text{if } \nu > \nu_\mathrm{p} \end{cases} \tag{5.2}$$

where ν_p is the *break frequency* at which the index changes from α to β, hence the name 'broken power-law function'. The function N_ν represents the number of γ-rays observed with frequencies between ν and $\nu + \mathrm{d}\nu$, per unit area, per unit time. To convert N_ν to the flux density F_ν at a particular frequency ν, simply multiply by that frequency:

$$F_\nu(\nu) = \nu N_\nu(\nu) \tag{5.3}$$

Example 5.1

Consider the high-energy ($\nu > \nu_{\rm p}$) segment of the GRB spectral energy distribution shown in Figure 5.3.

(a) Using the figure, estimate the slope of the segment as it is plotted, i.e. on logarithmic axes.

(b) Using your answer from part (a), calculate an approximate value for the spectral index β that is used in Equation 5.2.

Solution

(a) To solve this part we will estimate the gradient of the best-fitting model line. We start by picking two points and estimating their x (or E) and y (or νF_ν) values. To keep things as simple as possible, we will pick the points that have $\log(\nu F_\nu) = -10$ and $\log(\nu F_\nu) = -11$. The corresponding $\log(E)$ values for these points are approximately $\log(E) \approx 0.8$ and $\log(E) \approx 3.3$, respectively. The gradient $\beta_{\log}$ of the line between these two points, in logarithmic space, is then:

$$\beta_{\log} \approx \frac{-10 - (-11)}{0.8 - 3.3} \approx -0.4$$

(b) Equation 5.2 describes the prompt GRB spectrum in units of N_ν versus ν. Equation 5.3 tells us how to convert from F_ν to N_ν and rearranging gives

$$N_\nu = \frac{F_\nu}{\nu}$$

Multiplying the numerator and the denominator by ν, we obtain the conversion between N_ν and νF_ν

$$N_\nu = \frac{\nu F_\nu}{\nu^2}$$

Rewriting Equation 5.2 for $\nu > \nu_{\rm p}$ in terms of νF_ν we find:

$$\nu F_\nu \propto \nu^2 \left(\frac{\nu}{\nu_{\rm p}}\right)^\beta \propto \nu_{\rm p}^{-\beta} \nu^{\beta+2}$$

Comparing this expression with the result of part (a) we can infer that:

$$\beta + 2 \approx -0.4$$

and so $\beta \approx -2.4$.

Typical parameters of prompt GRB spectra

The best-fitting values of parameters, $\nu_{\rm p}$, α and β vary from burst to burst. The distribution of the *break energy* $\epsilon_{\rm p} = h\nu_{\rm p}$ is centred around 150 keV but it spans a wide range between about 10 keV and 200 MeV.

The α and β distributions are centred around $\alpha \approx -1$ and $\beta \approx -2.2$, but there is quite a large scatter in both index values from burst to burst.

It is important to keep in mind that the broken power-law is not a physically motivated model. Rather, it is just a easy-to use function that approximates the spectral shapes of most observed prompt GRB spectra. However, later in this chapter you will learn how the values of the parameters α, β and $\nu_{\rm p}$ can be used to infer what physical processes are operating to generate the γ-ray emission in GRBs.

Classification of GRBs

The catalogues collected by *CGRO*, *Swift* and *Fermi* contain measured properties for thousands of GRBs. By analysing these catalogues, astronomers have found that GRBs can be divided into two distinct categories based primarily on their durations.

Figure 5.4 shows the distributions of T_{90} for GRBs that were observed by *CGRO* BATSE, *Swift* BAT and *Fermi* GBM. All three distributions show a large population of GRBs with durations longer than 2 seconds: these are referred to as **long GRBs**. As well as the long GRB peak, all three distributions in Figure 5.4 also reveal a smaller population with durations shorter than 2 seconds: these are called **short GRBs**.

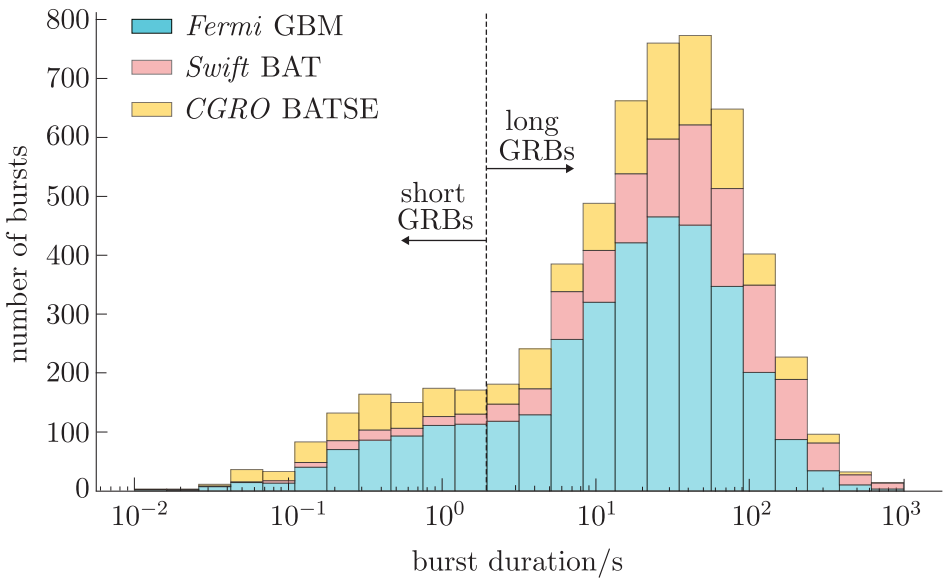


Figure 5.4 The distribution of burst durations (T_{90}), showing contributions for bursts detected by *CGRO* BATSE (yellow), *Swift* BAT (pink) and *Fermi* GBM (blue). The combined distribution is bimodal (i.e. it has two distinct peaks), showing two populations of bursts with typical durations that are either longer or shorter than 2 seconds. The peak for short GRBs is much less pronounced than the long GRB peak.

The distributions shown in Figure 5.4 show that astronomers have detected many more long GRBs than short GRBs. Figure 5.5 shows that short GRBs tend to appear fainter than long GRBs.

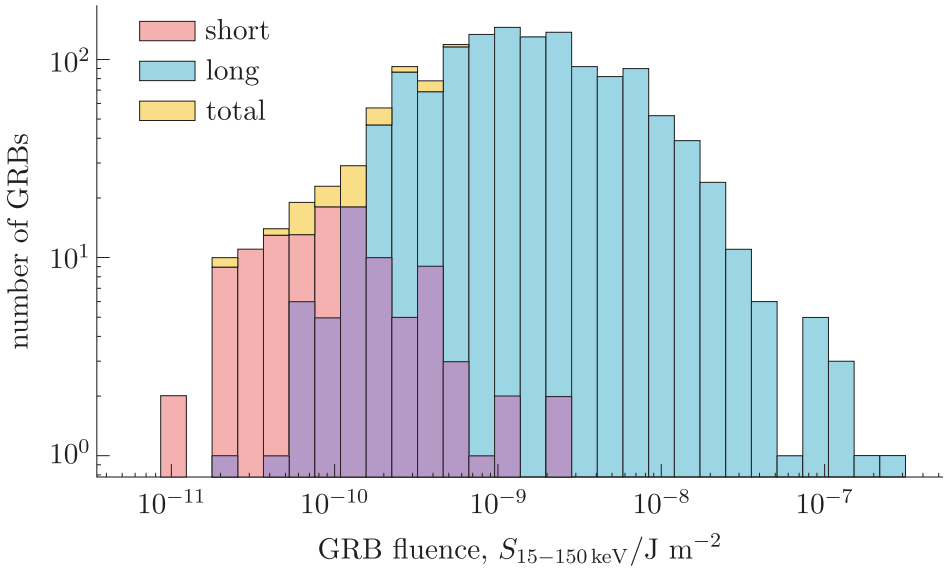


Figure 5.5 Distribution of prompt γ-ray fluences for GRBs observed by the *Swift* BAT instrument. The purple region shows the overlap between the distributions for long and short GRBs.

The populations of long and short GRBs are also observed to have different spectral properties. Although there is some overlap between the two GRB classes, a typical short GRB is likely to emit more of its flux as high-energy photons than a typical long GRB.

Later in the chapter you will see how observable differences between these two types of burst can be used to identify different populations of celestial objects that could be their progenitors.

Spatial distribution of GRBs

Figure 5.1 shows the locations of all GRBs observed by *CGRO* BATSE, *Swift* BAT and *Fermi* GBM. For all three instruments, the observed distribution of GRBs on the sky is isotropic, with no clustering of events in any direction. The distribution of GRB fluences is also highly isotropic with bursts of all brightnesses equally likely to appear in all directions.

Now let's consider the distribution of GRB distances. In 1997, GRB 970228 became the first GRB to have a reliably measured cosmological redshift; we will discuss this further in Section 5.2.2. Since then astronomers have measured the redshifts of hundreds of GRBs and have firmly established that they occur outside the Milky Way, in distant galaxies. Figure 5.6 shows the distribution of redshifts for 388 GRBs that were detected by the *Swift* space telescope. The most distant GRB in this sample is GRB 090423, which has a measured redshift of $z = 8.2$. GRB 090423 was observed in 2009 but its γ-rays were emitted when the Universe was just 617 million years old. Note that the short GRBs detected by *Swift* are fewer in number than the long GRBs at any redshift and the distribution of short GRBs is skewed towards much lower redshifts.

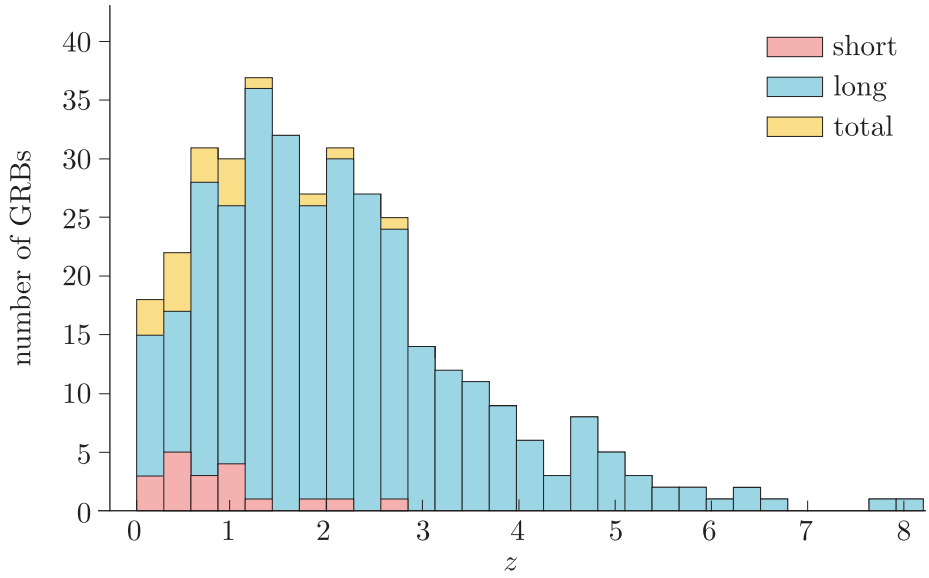


Figure 5.6 The distribution of redshifts (z) for 388 GRBs that were detected by the *Swift* space telescope.

Swift measured $S_{15-150\,\mathrm{keV}}$ for GRB 090423 to be $5.9 \times 10^{-10}\,\mathrm{J\,m^{-2}}$, which is actually relatively faint compared to the majority of bursts. Nonetheless, if its emission was isotropic its distance means that the total energy output of its prompt γ-ray emission could still have been as high as 3.2×10^{44} J. The fact that GRBs appear so bright in γ-rays, even though they occur at cosmological distances, confirms that they must be among the most energetic events in the Universe.

5.2.2 The GRB afterglow

The prompt γ-ray emission that we discussed in the last section is very conspicuous, which makes it relatively easy to detect GRBs when they happen. However, the prompt emission fades quickly and this makes it technically very challenging to search for and localise the celestial objects that produce the γ-rays. Historically, the task was made even harder, because many γ-ray telescopes had relatively coarse angular resolution. For example, the *CGRO* BATSE instrument could only detect γ-ray sources with a positional accuracy of about 4 degrees.

Fortunately, we now know that the brief, prompt emission from most GRBs is followed by a longer period of much fainter emission at lower frequencies. This lower-frequency emission is called the GRB **afterglow** and its extended duration gives astronomers the time they need to accurately determine GRB positions. Figure 5.7 shows the first GRB afterglow that was ever detected. It is the afterglow of GRB 970228, which was imaged in visible light by the William Herschel Telescope on the island of La Palma. The afterglow coincided with the outskirts of a distant galaxy at redshift $z = 0.695$, which allowed astronomers to finally establish that GRBs were extragalactic phenomena that were happening at cosmological distances.

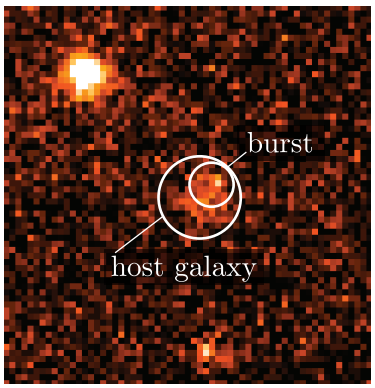


Figure 5.7 The optical afterglow of GRB 970228 superimposed on its host galaxy. The redshift of the host was later determined to be $z = 0.695$.

One of the main design goals of the *Swift* mission was the efficient detection of GRB afterglows. Almost all GRBs are followed by an X-ray afterglow, so the *Swift* satellite carries an X-ray telescope, called the XRT. When the BAT detects a GRB, *Swift* is designed to quickly reorient itself to point the XRT in the direction of the burst and start observing the afterglow. Roughly half of GRBs also exhibit a visible (optical) afterglow, so *Swift* also carries a telescope called UVOT that is sensitive to optical and ultraviolet photons.

X-ray afterglows

Figure 5.8 shows four typical X-ray afterglow light curves that were measured by the *Swift* XRT together with corresponding prompt emission measured by the *Swift* BAT. The afterglow light curves are typically much less chaotic than those you saw for the prompt emission in Figure 5.2, and they can normally be well modelled as a series of distinct segments forming a multiply-broken power law.

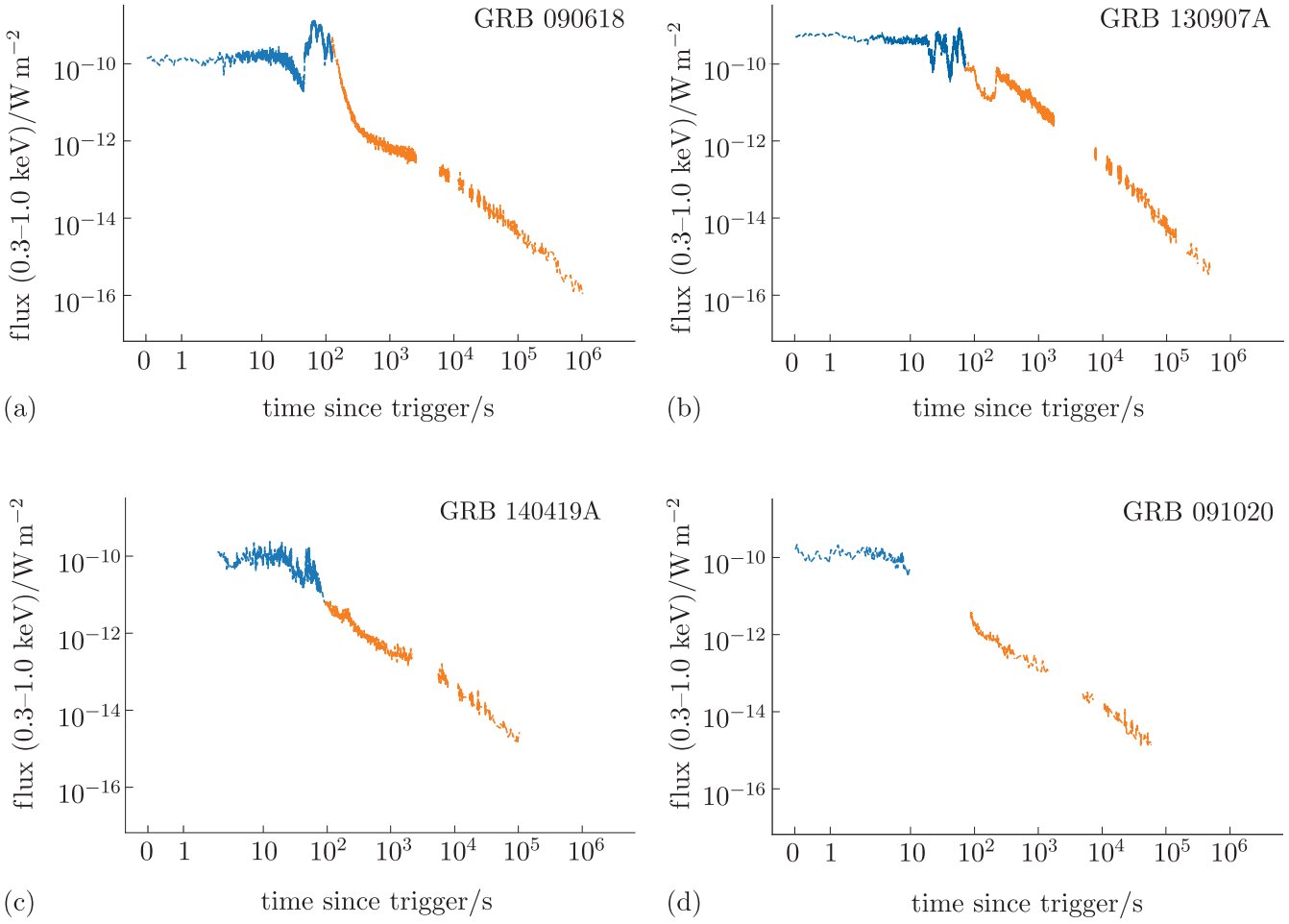


Figure 5.8 Four examples of GRB light curves that were measured by instruments on board the *Swift* satellite. The light curves show both the prompt emission measured by the BAT (blue) and the X-ray afterglow measured by the XRT (orange).

Within each segment, the flux $F_\nu(t)$ usually decreases over time following a power-law trend. Using the symbol κ to represent the power-law index within a particular segment, we can write

$$F_\nu(t) \propto t^{-\kappa}$$

where κ is a small positive number, but not in general an integer.

Analysis of early observations led astronomers to form a 'canonical' model for the X-ray light curve of GRB afterglows, with four distinct phases:

(a) The first phase is often called the *steep decay* phase because it is characterised by a rapid drop in flux over time. During this phase the prompt emission fades rapidly and the afterglow starts to dominate the X-ray flux. Observed values of the index κ during the steep decay phase span a range $1 \lesssim \kappa \lesssim 4$.

(b) The second phase is often referred to as the *plateau*. During the plateau phase the index $\kappa \sim 1$. The observed flux decreases much more slowly over time and can even start to increase slightly.

(c) The third phase is called the *normal decay* phase. In this phase, the flux begins to fall more steeply again but with observed values of κ that are smaller than they were in the steep decay phase.

(d) The final phase in this canonical model is called the *post jet-break* phase. We will discuss the reasons for this name later in the chapter. In this phase, the rate of flux decay steepens again with κ taking values between those of the steep decay and normal decay phases.

Many afterglow light curves also include short flares, when the flux increases rapidly for a short time before returning to its original power-law decay trajectory.

Thanks to hundreds of afterglow observations by the *Swift* XRT, astronomers now know that fewer than half of all GRBs exhibit all four canonical phases. Nonetheless, the canonical model remains useful for describing different behaviours we may observe as afterglows evolve and fade over time. Later in this chapter we will use it to help understand the processes that produce the afterglow emission; this will also give us clues about what could be driving and powering the prompt emission.

- Look again at the GRB afterglow light curves in Figure 5.8 and try to identify which, if any, of the canonical phases each one exhibits.
- Deciding whether a light curve exhibits a particular phase can be somewhat subjective so your answer may not exactly match those below.

  (a) Evidence of the steep decay phase is shown by GRB 090618 (between 100 s and 300 s after trigger) and GRB 140419A (between 70 s and 100 s after trigger).

  (b) Evidence of the plateau phase is shown by GRB 090618 (between 300 s and 1200 s after trigger) and possibly by GRB 130907A (around 100 s after trigger).

  (c) All four light curves seem to show a normal decay phase.

  (d) None of the light curves shows strong evidence for an obvious post jet-break phase.

The X-ray spectra during the afterglow are also well modelled by power-law functions. Observations by the *Swift* XRT have shown that once the prompt emission has faded, the afterglow spectral index remains quite stable throughout all canonical phases for most individual bursts. As a population, the majority of bursts have afterglow spectral indices in the range between 0.5 and 1.5.

Optical afterglows and achromatic breaks

Roughly half of the GRBs detected by *Swift* also emit optical afterglows that can accompany their X-ray counterparts but often outlast them. Indeed, optical afterglows can persist for several months after the prompt γ-ray emission has faded.

Figure 5.9 shows the optical and ultraviolet afterglow light curves of GRB 090618 that were observed in six different wavelength bands by the *Swift* UVOT instrument. Just like the X-ray afterglows, these optical light curves can be modelled using broken power laws. An interesting feature of these plots is that they all change index by roughly the same amount, at roughly the same time. Astronomers use the term **achromatic** to describe breaks like this, which exhibit the same behaviour across a wide range of photon frequencies. Not all GRB afterglow light curves exhibit achromatic breaks: later in this chapter you will see that they can be interpreted as a geometric effect that depends on the observer's viewing angle rather than some intrinsic change in the photon production mechanism.

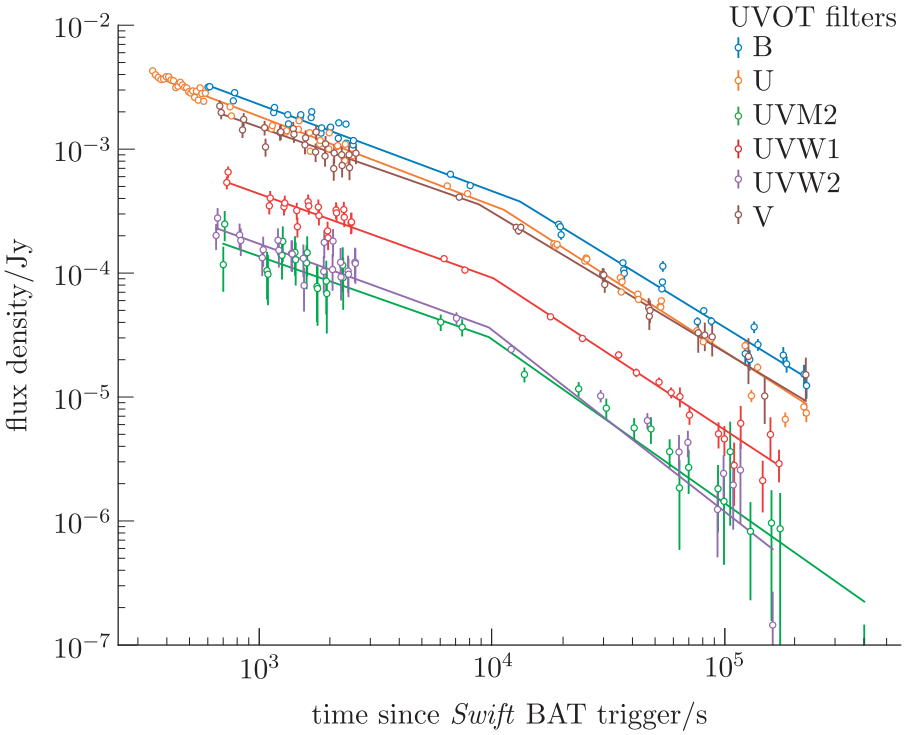


Figure 5.9 An achromatic break in the optical and ultraviolet afterglow light curves of GRB 090618 measured using the *Swift* UVOT. Each light curve was measured through a different optical or ultraviolet filter. All six exhibit a power-law break around 10^4 s after the prompt burst trigger.

5.2.3 GRB hosts and nurseries

The large number of optical and X-ray afterglows that have been detected by *Swift* and other telescopes has allowed astronomers to determine the properties of many GRBs' host galaxies and the locations of the GRBs within them. This has revealed a clear difference between the populations of galaxies that host long versus short GRBs. Long GRBs tend to be found within the bright star-forming regions (nurseries) of irregular or disturbed spiral galaxies. Long GRB hosts also tend to contain stars and interstellar media with low metallicities. Conversely, short GRBs are found in galaxies of all types, with no particular prevalence within star-forming regions. In Section 5.5 you will see that these differences between the environments where long and short GRBs occur can provide clues about the celestial objects that produce them.

5.3 From observations towards models

In this section we will examine how the observable properties you learned about in Section 5.2 can be used to infer a set of *physical* properties that must be realised by any plausible GRB progenitor candidates. In Section 5.5 we will use these physical properties as selection criteria to identify populations of celestial objects that could be the sources of GRBs.

5.3.1 Evidence for relativistic expansion

In this section we will examine several aspects of the observational evidence that you saw in Section 5.2. From these we can infer that:

(a) The physical processes that generate the prompt GRB emission must occur within a very compact region of space.

(b) The number density of γ-ray photons in these regions must be extremely large.

(c) The γ-ray-emitting material must be expanding ultra-relativistically.

Compactness

Example 5.2 shows how the variability of the prompt-emission light curves of GRBs can be used to infer an upper limit on the sizes of GRB progenitors. Remarkably, this limit turns out to be much smaller than the radius of a typical star like our Sun!

Example 5.2

Figure 5.2 shows that the prompt emission from GRBs can vary by large factors on millisecond (ms) timescales.

(a) Use this information to estimate an upper bound on the size of the region producing the prompt γ-rays.

(b) Compare this size estimate to the solar radius, $R_\odot \approx 7 \times 10^5$ km.

Solution

(a) We start by assuming that large changes in the observed flux must reflect changing physical properties across all or most of the γ-ray-emitting region.

Even if the physical changes causing variations in γ-ray *emission* happen instantaneously throughout the emitting region, the *observed* flux must vary more slowly because γ-rays emitted from different parts of the region must travel different distances to reach us. The different light travel times smooth out the *observed* flux variability. This means that the minimum observed variability timescale corresponds roughly with the typical time taken for light to propagate across the γ-ray-emitting region.

If we use Δt_{min} to denote the minimum observed variability timescale and l to denote the size of the γ-ray-emitting region along the line of sight, then we can write:

$$\Delta t_{\text{min}} \gtrsim \frac{l}{c}$$

Rearranging this equation and using $\Delta t_{\text{min}} \approx 10^{-3}\,\text{s}$ we find that

$$\begin{aligned} l &\lesssim \Delta t_{\text{min}} c \\ &\lesssim 10^{-3}\,\text{s} \times 3 \times 10^{8}\,\text{m}\,\text{s}^{-1} \\ &\lesssim 3 \times 10^{5}\,\text{m} \end{aligned} \tag{5.4}$$

(b) We have estimated that the region producing the prompt GRB photons is around 300 km in size, which is roughly 0.05% of the solar radius!

The result of Example 5.2 tells us that the prompt emission from GRBs must be produced within remarkably compact regions of space. Next, let's consider what our observations imply about the physical conditions within those regions, starting with an estimate of the γ-ray photon number density.

Photon number density

Before we can estimate the number density of γ-ray photons within the region that produces the prompt GRB emission, we must first estimate the typical γ-ray luminosities of long and short GRBs. The typical 15–150 keV fluence of a long GRB is $S_{15\text{–}150\,\text{keV}} \sim 10^{-8}\,\text{J}\,\text{m}^{-2}$ and Figure 5.4 shows the typical duration of a long GRB is $T_{90} \sim 80\,\text{s}$. For a typical long GRB with redshift $z \sim 2$, and assuming that GRBs emit isotropically, this implies a long GRB luminosity $L_{15\text{–}150\,\text{keV}} \sim 4 \times 10^{42}\,\text{W}$, which is about 10^{16} times more luminous than the Sun!

The following exercise asks you to calculate a luminosity estimate for short GRBs. You should find that they are even brighter.

Exercise 5.1

Assuming that short GRBs emit isotropically, use the values for their typical fluence and duration shown in Figures 5.5 and 5.4 to estimate the 15–150 keV luminosity of their prompt emission in watts. You may assume that the typical observed redshift for short GRBs, shown in Figure 5.6, equates to a luminosity distance $d_{\mathrm{L}} \sim 6200\,\mathrm{Mpc}$.

We can use our GRB luminosity estimates to calculate the number density of photons within the prompt γ-ray emission region.

Assuming a spherical emitting region, the flux of γ-ray photons escaping through a small area element on the surface of the emitting region, ΔA, is given by $F = \Delta E/(\Delta t\,\Delta A)$, where ΔE is the energy of the photons that pass through ΔA in a time interval Δt. The flux is therefore given by

$$F = \frac{\epsilon_\gamma n_\gamma \Delta V}{\Delta t\,\Delta A} \tag{5.5}$$

where n_γ is the photon number density in the emitting region, and ϵ_γ is the typical photon energy. The photons propagate at the speed of light, and so

$$\Delta V = c\,\Delta t\,\Delta A$$

Therefore the flux can be expressed as

$$F = c n_\gamma \epsilon_\gamma \tag{5.6}$$

Assuming that the GRBs emit isotropically, we can also write F in terms of the GRB γ-ray luminosity L and the surface area of the emission region of radius R:

$$F = \frac{L}{4\pi R^2}$$

By combining these two expressions for F and rearranging we can derive an expression for n_γ in terms of L.

$$n_\gamma = \frac{L}{4\pi R^2 c \epsilon_\gamma} \tag{5.7}$$

The following example uses all of the results that we have derived so far in this section to estimate some numerical values for n_γ in GRBs.

Example 5.3

Figure 5.3 shows that prompt GRB spectra are dominated by γ-ray photons with energies $\sim$500 keV; the observed variability of GRB light curves tells us that those photons are generated within regions that are at most a few hundred kilometres in size. Use this information to calculate an order-of-magnitude estimate for the γ-ray photon number density within the emission region of a GRB with luminosity $L_\gamma = 10^{44}\,\mathrm{W}$. Give your answer in SI units.

Solution

To solve this problem we can use Equation 5.7. However, the question asks for an answer in SI units so we must first convert the γ-ray energy from keV to joules. The conversion factor between electronvolts and joules is just the electron charge $e = 1.602 \times 10^{-19}$ C. This means that

$$500\,\text{keV} = 500 \times 1000 \times 1.602 \times 10^{-19}\,\text{J} = 8.01 \times 10^{-14}\,\text{J}$$

Now we just need to evaluate Equation 5.7. We only need to calculate an order-of-magnitude estimate, so will assume that all the γ-rays have $\epsilon_\gamma = 10^{-13}$ J and that the average radius of the prompt-emission region is $R = 10^5$ m. Using these values

$$\begin{aligned} n_\gamma &= \frac{L_\gamma}{4\pi R^2 c \epsilon_\gamma} = \frac{10^{44}\,\text{W}}{4\pi \times (10^5\,\text{m})^2 \times 3 \times 10^8\,\text{m s}^{-1} \times 10^{-13}\,\text{J}} \\ &= 2.7 \times 10^{37}\,\text{m}^{-3} \approx 10^{37}\,\text{m}^{-3} \end{aligned}$$

Using the approach demonstrated in Example 5.3, it is found that that typical GRB luminosities in the range $\sim 10^{42}$–10^{44} W imply values of n_γ in the range $n_\gamma \sim 10^{35}$–$10^{37}\,\text{m}^{-3}$. In the next section you will see that these huge photon number densities provide very strong evidence for relativistic expansion in GRBs.

The compactness problem

If their energies are high enough, γ-ray photons that collide with each other undergo electron–positron pair production (see *Cosmology* Chapter 8). For this to happen, the energies of the two photons must be *at least* equal to the combined rest mass $2m_\text{e}$ of the electron and positron, in a frame of reference in which their combined momentum is zero.[†] Equation 5.8 expresses this criterion

$$\epsilon_1 \epsilon_2 \gtrsim (m_\text{e} c^2)^2 \tag{5.8}$$

We saw in Section 5.2.1 that the *observed* spectra of GRBs exhibit significant fluxes of photons with energies well above the rest-mass energy of the electron. This means the vast majority of photon pairs that collide within the prompt-emission region will have enough energy to produce an electron–positron pair.

The $e^\mp$ pair production cross-section

If two γ-ray photons with sufficient energy interact, they are not guaranteed to annihilate and produce an $e^\mp$ pair. The probability that they do annihilate is quantified by the $e^\mp$ pair production cross-section, which we will denote $\sigma_{\gamma\gamma}$ throughout this chapter.

[†]You may see this frame referred to as the centre-of-momentum frame.

The exact value of $\sigma_{\gamma\gamma}$ depends on the product of the interacting photons' energies $\epsilon_1\epsilon_2$ and the angle between their trajectories $\theta_{\gamma\gamma}$. However, it is often sufficient to consider an average value and use the approximation $\sigma_{\gamma\gamma} \approx \sigma_{\mathrm{T}}/4$, where σ_{T} is the Thomson cross-section.

The probability that a γ-ray will escape from the prompt-emission region of a GRB without being annihilated can be described in terms of a quantity called the **optical depth**, which we will denote $\tau_{\gamma\gamma}$. For the $e^{\mp}$ pair production process, $\tau_{\gamma\gamma}$ can be written approximately in terms of the γ-ray number density n_γ, the cross-section for the pair production process $\sigma_{\gamma\gamma}$ and the distance Δl that a typical γ-ray must travel to escape the prompt emission. To derive rough estimates for $\tau_{\gamma\gamma}$ in GRBs, we can assume that Δl approximately equals the size of the emission region R and write:

$$\tau_{\gamma\gamma} \approx \sigma_{\gamma\gamma} n_\gamma \, \Delta l \sim \sigma_{\gamma\gamma} n_\gamma \, R \tag{5.9}$$

The probability that a photon escapes a region with optical depth $\tau_{\gamma\gamma}$ without being annihilated is $e^{-\tau_{\gamma\gamma}}$. Using the values of $R \sim 100\,\mathrm{km}$ and $n_\gamma \sim 10^{37}\,\mathrm{m}^{-3}$ that we estimated earlier in the chapter, and assuming $\sigma_{\gamma\gamma} \approx \sigma_{\mathrm{T}}/4$, we calculate $\tau_{\gamma\gamma} \sim 10^{13}$–$10^{14}$. Such an enormous optical depth means that it is almost inevitable that all γ-rays in the prompt-emission region will undergo pair production before they escape, which introduces an apparent contradiction.

We have shown that almost all the γ-rays produced within the prompt-emission region (and particularly those with energies greater than $\sim$0.5 MeV) should undergo pair production and be converted into electrons and positrons before they escape – we should never observe them. However, we *do* observe γ-rays with energies much larger than 0.5 MeV in the spectra of GRBs: TeV photons have been observed from some of them! This contradiction is called the *compactness problem* and for many years it was used to argue that GRBs could not be very luminous and therefore could not be at cosmological distances.

The effect of relativistic expansion

Ultimately, it was realised that the compactness problem can be resolved if the γ-ray-emitting material in the GRB prompt-emission region moves relativistically. To see why, we will first consider the effects of bulk relativistic motion on the observed properties of GRB photons. Properly accounting for those effects will help to explain how the γ-rays we detect avoid annihilation and escape the prompt-emission region of the GRB.

We will consider a scenario in which the prompt-emission region and the material that it contains is moving towards us with a bulk Lorentz factor $\gamma \gg 1$. Unless stated otherwise, the symbol γ throughout this chapter refers to the *bulk* Lorentz factor of a region containing γ-ray-emitting material. We will also use the symbols S and S$'$ to represent the observer's rest frame and the rest frame of the prompt emission, respectively (analogous to our consideration of relativistic speeds in Chapter 4).

Accounting for relativistic blueshifting

Highly relativistic motion means that any prompt photons we detect from the GRB will have been strongly blueshifted.

- If we observe a prompt γ-ray from our GRB with energy ϵ_γ in S, what energy would we measure for the same photon in S′?
- The relativistic Doppler shift formula (Equation 4.11) tells us that we would measure an energy $\epsilon'_\gamma = \mathcal{D}^{-1}\epsilon_\gamma$ (since $\epsilon_\gamma = h\nu$).

For $\gamma > 1$, the energy of the γ-rays in S′ is less than the energy we observe.

If we do not account for relativistic effects, then we overestimate the number of γ-rays in the prompt-emission region that have enough energy to undergo electron–positron pair production, and therefore we overestimate the optical depth to pair production and how difficult it is for photons to escape.

Let's estimate the factor by which we overestimate the number of photon pairs that can undergo $e^{\mp}$ pair production when we ignore relativistic effects. To do this, we will first find an expression for the number of photons produced during the burst (i.e. within the variability timescale Δt) that exceed an arbitrary energy threshold, $\epsilon_{\gamma,\mathrm{t}}$. We will then use this expression to derive the number of γ-rays emitted by the GRB that have enough energy to undergo pair production if they interact with a photon that has energy $\epsilon_\gamma \geq \epsilon_\mathrm{t}$. We will find that this number depends on the bulk Lorentz factor of the γ-ray-emitting region and the observed spectrum of the γ-ray emission.

$N_\epsilon(\epsilon_\gamma)\,\mathrm{d}\epsilon_\gamma$ represents the number of γ-ray photons with observed energies between ϵ_γ and $\epsilon_\gamma + \mathrm{d}\epsilon_\gamma$, per unit area, per unit time.

Based on what you learned in Section 5.2.1 it makes sense to model the photon flux $N_\epsilon(\epsilon_\gamma)\,\mathrm{d}\epsilon_\gamma$ emitted in a time window of Δt for $\epsilon_\gamma > \epsilon_{\gamma,\mathrm{t}}$ as a falling power-law function. Using f_γ to represent a constant normalisation term and χ to represent a power-law index, we can write:

$$N_\epsilon(\epsilon_\gamma)\,\mathrm{d}\epsilon_\gamma = f_\gamma \epsilon_\gamma^{-\chi}\,\mathrm{d}\epsilon_\gamma \tag{5.10}$$

To find the flux of γ-rays with observed energies above ϵ_t we integrate Equation 5.10 with respect to photon energy:

$$\begin{aligned} N_\epsilon(\epsilon_\gamma > \epsilon_{\gamma,\mathrm{t}}) &= \int_{\epsilon_{\gamma,\mathrm{t}}}^{\infty} N(\epsilon_\gamma)\,\mathrm{d}\epsilon_\gamma = \int_{\epsilon_{\gamma,\mathrm{t}}}^{\infty} f_\gamma \epsilon_\gamma^{-\chi}\,\mathrm{d}\epsilon_\gamma \\ &= f_\gamma \left[\frac{\epsilon_\gamma^{1-\chi}}{1-\chi}\right]_{\epsilon_{\gamma,\mathrm{t}}}^{\infty} \\ &= f_\gamma \frac{\epsilon_{\gamma,\mathrm{t}}^{1-\chi}}{\chi - 1} \end{aligned} \tag{5.11}$$

Now, if d_L is the luminosity distance to the GRB, then the total *number* of photons with observed energies above $\epsilon_{\gamma,\mathrm{t}}$ that a GRB emits during the time interval Δt is:

$$N(\epsilon_\gamma > \epsilon_{\gamma,\mathrm{t}}) = 4\pi d_\mathrm{L}^2 \Delta t f_\gamma \frac{\epsilon_{\gamma,\mathrm{t}}^{1-\chi}}{\chi - 1} \tag{5.12}$$

We want to compare this with the number of γ-rays that would have energies $\epsilon'_\gamma > \epsilon_{\gamma,\mathrm{t}}$ measured in the *rest frame* of the prompt-emission region, S′. Note that by comparing numbers of photons during the burst, we do not need to transform the time interval between frames, because the total number of photons emitted is invariant.

If we assume that $\gamma \gg 1$, then the relativistic Doppler shift formula (Equation 4.11) can be approximated as $\epsilon'_\gamma \approx \epsilon_\gamma/2\gamma$ and therefore:

$$\begin{aligned} N(\epsilon'_\gamma > \epsilon_{\gamma,\mathrm{t}}) &= N(\epsilon_\gamma/2\gamma > \epsilon_{\gamma,\mathrm{t}}) = N(\epsilon_\gamma > 2\gamma\epsilon_{\gamma,\mathrm{t}}) \\ &= 4\pi d_\mathrm{L}^2 \Delta t f_\gamma \frac{(2\gamma\epsilon_{\gamma,\mathrm{t}})^{1-\chi}}{\chi - 1} \\ &\propto \gamma^{1-\chi} N(\epsilon_\gamma > \epsilon_{\gamma,\mathrm{t}}) \propto \frac{1}{\gamma^{\chi-1}} N(\epsilon_\gamma > \epsilon_{\gamma,\mathrm{t}}) \end{aligned} \tag{5.13}$$

Equation 5.13 tells us $N(\epsilon'_\gamma > \epsilon_{\gamma,\mathrm{t}}) \propto \gamma^{1-\chi}$ and that we would overestimate the number of photons that can undergo $\mathrm{e}^\mp$ pair production by a factor $\gamma^{\chi-1}$ if we ignore relativistic effects.

For pair production to occur, two photons need to independently exceed a threshold energy, such that their combined energy satisfies Equation 5.8. For example, consider the population of photons with prompt-emission region rest-frame energies $\epsilon'_{\gamma,1} \gtrsim \epsilon_{\gamma,\mathrm{t}}$. There are a factor $\gamma^{\chi-1}$ fewer of these photons available in S′ than we would infer from the observed spectrum without accounting for the relativistic blueshift.

Now, each of the photons that *do* have energies $\epsilon'_{\gamma,1}$ exceeding $\epsilon_{\gamma,\mathrm{t}}$ can only undergo pair production if they encounter another photon with energy $\epsilon'_{\gamma,2} \gtrsim 2(m_\mathrm{e}c^2)^2\epsilon_{\gamma,\mathrm{t}}^{-1}$ and the number of photons that fulfil *this* criterion is *also* a factor $\gamma^{\chi-1}$ smaller than we would infer without taking blueshifting into account. Overall, the total number of pairs that can undergo pair production is reduced by two factors of $\gamma^{\chi-1}$, resulting in a total reduction by a factor of approximately $\left(\gamma^{\chi-1}\right)^2 = \gamma^{2\chi-2}$.

Accounting for timescales in the relativistic outflow

As well as making us overestimate the number of photon pairs that can undergo pair production, the relativistic speed of the outflow also means that we underestimate the size of the prompt-emission region.

Consider an observer who detects two photons that arrive from a region that is moving towards them with a highly relativistic speed, such that the region's bulk Lorentz factor $\gamma \gg 1$. This situation has some similarities with the superluminal motion discussion in Chapter 4 of this book, where the second photon has less distance to travel than the first, because of the movement of the emitting region towards us in the intervening time.

In fact, it can be shown that

$$\Delta t_\mathrm{obs} \approx \frac{\Delta t_\mathrm{em}}{2\gamma^2} \tag{5.14}$$

This can be seen by noting that the factor relating the time intervals in Example 4.2 is equivalent to $\gamma\mathcal{D}$, and then using the approximation for $\mathcal{D}$ introduced in the previous section.

The timescales of processes in the observer's frame therefore appear shorter by a factor $\sim\gamma^2$ if they occur in regions that are moving towards the observer at relativistic speeds. Consequently, if we want to use the observed minimum variability timescale $\Delta t_{\rm min}$ to estimate the size of the prompt-emission region, the correct expression is not $R \sim c\,\Delta t_{\rm min}$, but rather:

$$R \sim \gamma^2 c\,\Delta t_{\rm min} \tag{5.15}$$

In other words, the prompt-emission region is a factor $\sim\gamma^2$ larger than we would estimate if we ignore the effects of relativistic motion of the emitting region towards the observer during the time interval considered.

Resolving the compactness problem

To see how accounting for relativistic bulk motion can resolve the GRB compactness problem, let's derive a rough expression that describes how the $e^{\mp}$ pair production optical depth $\tau_{\gamma\gamma}$ varies as a function γ.

Earlier, we showed that the number of γ-ray pairs in the prompt-emission region that can undergo $e^{\mp}$ pair production is proportional to $\gamma^{2-2\chi}$. This means that for any randomly chosen γ-ray the number density of γ-rays in the prompt-emission region with which it can undergo pair production also changes by a factor of $\gamma^{2-2\chi}$. The number density n_γ is also inversely proportional to the volume of the prompt-emission region. Assuming that the volume is roughly proportional to R^3 using Equation 5.15, we can say:

$$n_\gamma \propto \frac{\gamma^{2-2\chi}}{R^3} \propto \gamma^{2-2\chi}\gamma^{-6}$$

As we did when deriving Equation 5.9, we can assume that the distance that γ-rays propagate before they escape the prompt-emission region is roughly R and so using Equation 5.9, we can write:

$$\tau_{\gamma\gamma} \propto n_\gamma R \propto \gamma^{2-2\chi}\gamma^{-6}\gamma^2 \propto \gamma^{2-2\chi}\gamma^{-4} \propto \gamma^{-2\chi-2}$$

This result tells us that if we set $\gamma \approx 1$ and ignore the effects of relativistic motion we would overestimate the optical depth in the prompt-emission region by a factor $\propto \gamma^{2\chi+2}$.

Our derivation has made several approximations and assumptions related to aspects of our model like the observed γ-ray spectrum and the geometry of the prompt-emission region. Different assumptions would yield slightly different relationships between τ and γ. However, it would always be the case that accounting for bulk relativistic motion of the prompt-emission region implied much lower optical depths for $e^{\mp}$ pair production.

For a typical GRB with an observed minimum variability timescale $\sim 10^{-3}$ s, an observed γ-ray luminosity $\sim 10^{44}$ W and a high-energy spectral index $\chi \sim 2$, $\tau_{\gamma\gamma}$ falls below 1 and the compactness problem is avoided if $\gamma \gtrsim 100$. For context, this value is ten times larger than the bulk Lorentz factors that have been inferred for AGN jets! Based on our solution to the compactness problem, we have found that GRBs must produce the most highly relativistic outflows in the Universe!

Exercise 5.2

Consider a GRB with an observed minimum variability timescale $\Delta t_{\text{min}} \approx 1$ ms. Assume that the prompt-emission region of the GRB is moving towards us with a bulk Lorentz factor $\gamma \sim 100$.

(a) Estimate the size of the prompt-emission region R, in kilometres, accounting properly for its relativistic motion.

(b) Compare this estimated size with the radius of the Sun.

5.4 The fireball model

In the previous section we used observed prompt spectra and light curves to infer that GRBs must generate huge γ-ray luminosities $\sim 10^{44}$ W, on short timescales and within compact regions of space. We also showed that GRBs must generate γ-ray-emitting material that moves relativistically with spectacular bulk Lorentz factors $\gamma \sim 100$.

In this section we will introduce a model called the **fireball model** that can reproduce these physical properties, and can also explain how the GRB emission evolves from the prompt-emission phase and into the afterglow. The fireball model makes no assumptions about the sources of matter and energy in GRBs. Instead, all of the processes that are described by the fireball model are assumed to be driven by a hidden energy source that is usually referred to as the **central engine** of the GRB. In Section 5.5 we will consider two families of celestial objects that are considered to be plausible central engine candidates.

Figure 5.10 illustrates how the observed behaviours of GRBs are explained as arising from different stages in the evolution of the fireball model. We will now examine each of these stages in more detail. We start in the next section by considering the first stage of the fireball model, which predicts rapid initial expansion at highly relativistic speeds.

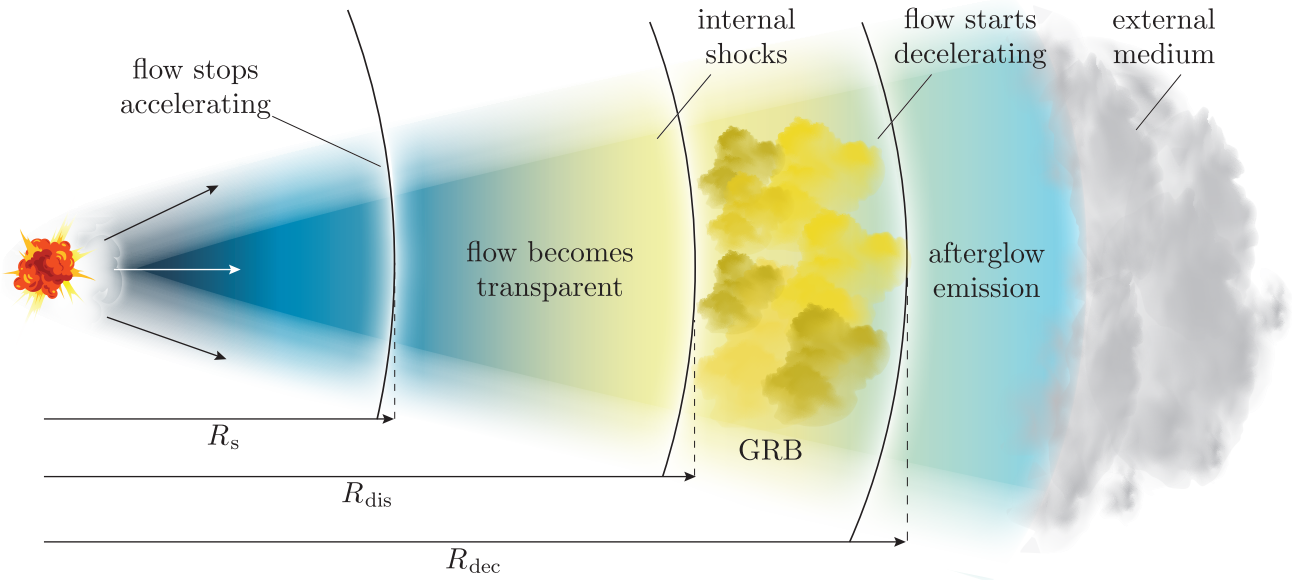


Figure 5.10 Schematic illustration of the fireball model for GRB emission. The diagram shows how different radiative processes happen at different distances from the exploding central engine.

5.4.1 Relativistic expansion

The initial conditions of the fireball model assume that a very large amount of energy E_0 is imparted instantaneously to matter with mass M_0 such that $E_0 \gg M_0c^2$ within a compact region of space that has a radius R_0. The result is a dense mixture of high-energy γ-ray photons and an ionised plasma containing electrons and positrons as well as baryonic particles. At this stage, the rate of Thomson and Compton scattering interactions between photons and electrons is so high that the fireball is effectively opaque, the γ-rays are trapped and the resultant radiation pressure is enormous.

The fireball remains effectively opaque to γ-ray photons as long as the rate of Compton scattering within it remains high. This opacity results in a very high γ-ray radiation pressure that must drive a fast expansion into the surrounding space.

We will model the expanding fireball as a spherical shell that has an inner radius R and an outer radius $R + \Delta R$. As the fireball expands, its internal energy is converted into bulk radial motion of the baryons and electrons it contains. If we assume that the expansion is adiabatic, so that no energy enters or leaves the fireball, and that radiation pressure drives the expansion, then it can be shown that the shell's bulk Lorentz factor grows in proportion with its inner radius R until it reaches a maximum value γ_{max}. Writing γ mathematically as a function of R:

$$\gamma(R) \propto R \tag{5.16}$$

Our assumption of adiabatic expansion means that we can estimate the maximum Lorentz factor that the fireball shell attains. To simplify our

derivation, let's also assume that the electrons and baryons in the fireball are initially at rest, with bulk Lorentz factor $\gamma_0 = 1$. This means that the total energy of the fireball when $R = R_0$, at the moment it starts expanding, is just

$$E_{\text{init}} = M_0 c^2 + E_0$$

where M_0 is the total mass contained in the fireball shell and E_0 is the contribution of energy from the central engine, whatever it is. Note that we make no assumptions yet about what type of celestial object the central engine might be.

When the fireball attains its maximum Lorentz factor, the electrons and baryons are still present but all of the energy represented by E_0 has been converted into bulk kinetic energy. This means that we can use Einstein's formula for the total energy to write

$$E_{\text{final}} = \gamma_{\max} M_0 c^2$$

Now we recall that the expansion is adiabatic so we can just equate our expressions for E_{init} and E_{final} and rearrange to find $\gamma_{\max}$:

$$\begin{aligned} E_{\text{final}} &= E_{\text{init}} \\ \implies \gamma_{\max} M_0 c^2 &= M_0 c^2 + E_0 \end{aligned}$$

Dividing through by $M_0 c^2$ gives

$$\begin{aligned} \gamma_{\max} &= \frac{M_0 c^2 + E_0}{M_0 c^2} \\ &= 1 + \frac{E_0}{M_0 c^2} \approx \frac{E_0}{M_0 c^2} \end{aligned} \tag{5.17}$$

Now let's derive an expression for the inner radius of the shell when it reaches its maximum velocity. This radius is often called the **saturation radius** and we will denote it using the symbol R_{S}. Using Equation 5.16 and our assumption that the shell starts from rest, we can write:

$$\frac{\gamma_{\max}}{\gamma_0} = \frac{R_{\text{S}}}{R_0}$$

This implies that

$$\begin{aligned} R_{\text{S}} &= R_0 \gamma_{\max} \\ &= R_0 \left(1 + \frac{E_0}{M_0 c^2}\right) \approx R_0 \frac{E_0}{M_0 c^2} \end{aligned} \tag{5.18}$$

Exercise 5.3

Consider a GRB fireball that expands from an initial radius $R_0 = \text{R}_\odot$ and achieves a maximum bulk Lorentz factor $\gamma_{\max} = 200$.

(a) Calculate the saturation radius for this GRB.

(b) If the initial energy of the fireball $E_0 = 10^{44}\,\text{J}$, estimate the mass M_0 of its matter content in solar mass units.

Radiation pressure can only accelerate the electrons and baryons in the fireball while it remains opaque to γ-ray photons.[‡] The photons accelerate electrons and positrons by repeated Compton scattering; the electrons accelerate the baryons via electrostatic Coulomb interactions. However, the fireball must eventually become transparent to γ-rays or we would never observe the prompt GRB emission. In the next section we will describe the fireball model's predictions for how, when and where this transition to transparency occurs.

5.4.2 Transparency and baryon loading

From a theoretical perspective, the γ-ray spectrum that is predicted by the fireball model depends quite sensitively on when the expanding fireball plasma becomes transparent. In the model, the fireball starts expanding from a hot dense initial state, and repeated scattering interactions during the expansion phase are expected to leave the γ-ray photons with a black-body spectrum. However, you saw in Section 5.2.1 that the prompt spectra of GRBs are power-law functions of photon energy, which are highly non-thermal, i.e. far removed from the spectrum of a black body.

To resolve this apparent contradiction, we must assume that the γ-rays we observe are not the same ones that drove the initial expansion and therefore we deduce that those earlier γ-rays have transferred all of their energy into bulk motion of electrons and baryons. This implies that for the GRBs we observe, the fireball remains opaque until after it has attained its maximum velocity and only becomes transparent when its inner radius is larger than $R_{\rm S}$. In Section 5.4.3 we will discuss how the bulk kinetic energy of the electrons and baryons is converted back into γ-rays to produce the prompt-emission spectrum that we observe.

The overall mass of the fireball (M_0) is dominated by the baryons in the plasma, which are thousands of times heavier than the electrons and positrons. For this reason, M_0 is often referred to as the **baryon load** of the fireball. However, the opacity of the fireball depends primarily on the number density of electrons it contains. When this density falls below a critical threshold, the fireball becomes transparent and the γ-rays can escape. The electron density is directly related to M_0 because the numbers of electrons, positrons and baryons in the fireball plasma are roughly equal. Therefore, we can use arguments based on opacity to constrain the range of M_0 values that are possible. If M_0 was too small ($\lesssim 10^{-7}\,{\rm M}_\odot$) then the opacity would fall too quickly as the fireball expands and the density decreases. The fireball would become transparent before its inner radius reaches $R_{\rm S}$ and we would observe a thermal GRB spectrum. Conversely, if M_0 is too large ($\gtrsim 5\times10^{-5}\,{\rm M}_\odot$) then the fireball would not have enough energy to attain an ultra-relativistic bulk Lorentz factor, the

[‡]In the very early stages of the fireball expansion, its opacity is provided by electrons and positrons produced by the pair production process you learned about in Section 5.3.1. However, these pairs annihilate quickly to leave the opacity dominated by electrons that were present in the initial fireball plasma.

compactness problem would not be circumvented, and $e^{\mp}$ pair production would eliminate all γ-rays with energies $\gtrsim 0.5\,\text{MeV}$. The combined constraints limit M_0 quite specifically to be a small fraction of a solar mass. We can use this narrow range of values in conjunction with Equation 5.17 to derive additional constraints on the bulk Lorentz factors that are reached in GRBs.

Exercise 5.4

Consider the example of GRB 090423, which we estimated in Section 5.2.1 to have a total energy output $E_0 \sim 10^{44}\,\text{J}$. Using limits on M_0 derived above, estimate the corresponding range of possible maximum bulk Lorentz factors γ_{max} that could have been reached in GRB 090423.

5.4.3 Generating the prompt γ-ray emission

In the previous section we argued that the expanding fireball shell must remain opaque until all of the energy in photons has been used to accelerate the baryon load of the fireball to ultra-relativistic speeds.

When the fireball becomes transparent it is expanding ultra-relativistically, but in the rest frame of the expanding shell the baryons and electrons have approximately black-body energy distributions and their velocities are at most mildly relativistic. The radiation produced by these particles would exhibit a black-body spectrum, which does not match the observed power-law spectra of prompt GRB emission.

In this section we will revisit two physical mechanisms that you learned about in Chapter 4. The fireball model invokes these mechanisms to:

(a) Transfer the bulk kinetic energy of baryons into random kinetic energy of electrons in the expanding shell rest frame and leave them with a power-law distribution of ultra-relativistic energies.

(b) Convert this random electron motion into high-energy γ-rays, thereby producing the prompt GRB emission that we ultimately observe.

Internal shocks

In Chapter 4 you learned how charged particles can obtain a power-law distribution of energies via diffusive shock acceleration as they scatter repeatedly across shock fronts in the jets and radio lobes of AGNs. The fireball model invokes diffusive shock acceleration as a mechanism for converting the bulk kinetic energy of the baryons in the fireball into random high-speed motion of electrons and positrons in the plasma.

To generate the shocks, the fireball model assumes that the thick expanding shell is really a collection of thin shells that are launched at slightly different times and have slightly different expansion speeds. In this scenario, shocks occur when faster shells that were launched later catch up with and overtake slower ones that were launched earlier. These shocks are often referred to as **internal shocks** because they result from interactions

between different components of the fireball, rather than interactions between the fireball and an external medium.

The innermost radius of the fireball when the thin shells begin to overtake one another and the internal shocks are formed is called the **dissipation radius**, denoted by R_{dis}. A good approximation for R_{dis} can be expressed in terms of the fireball's maximum Lorentz factor γ_{max} and the observed variability timescale Δt

$$R_{\text{dis}} \approx 2\gamma_{\text{max}}^2 c\,\Delta t \tag{5.19}$$

The following example shows how this result can be used to estimate a numerical value for R_{dis} for a typical GRB bulk Lorentz factor and variability timescale.

Example 5.4

Consider a GRB that ejects thin spherical shells at times that are, on average, separated by an interval $\Delta t = 10\,\text{ms}$. The shells have a small range of ultra-relativistic, but slightly different, speeds.

(a) If the average shell speed $v_{\text{ave}} = 0.9999c$, estimate the corresponding average shell Lorentz factor, γ_{ave}.

(b) Hence, estimate the radius R_{dis} at which the faster shells would start to catch the slower ones and internal shocks could form. Write your answer in kilometres.

Solution

(a) To solve this part, we just use the formula for the Lorentz factor:

$$\gamma_{\text{ave}} = \left(1 - \frac{v_{\text{ave}}^2}{c^2}\right)^{-1/2} = \left(1 - 0.9999^2\right)^{-1/2}$$
$$\approx 70.71$$

(b) Now, using Equation 5.19:

$$R_{\text{dis}} \approx 2\gamma_{\text{ave}}^2 c\,\Delta t$$
$$\approx 2 \times (70.71)^2 \times 3 \times 10^8\,\text{m s}^{-1} \times 10^{-2}\,\text{s}$$
$$\approx 3 \times 10^7\,\text{km}$$

γ-rays from synchrotron radiation

Acceleration by internal shocks leaves electrons in the fireball with a negative-index power-law distribution of Lorentz factors. Using the symbol $p > 0$ to represent the power-law index, we can express the number of electrons with Lorentz factors between γ_{e} and $\gamma_{\text{e}} + \text{d}\gamma_{\text{e}}$ as:

$$N(\gamma_{\text{e}})\,\text{d}\gamma_{\text{e}} \propto \gamma_{\text{e}}^{-p}\,\text{d}\gamma_{\text{e}} \tag{5.20}$$

Note that γ_{e} refers to the Lorentz factors of individual electrons and not to the bulk Lorentz factor of the prompt-emission region, as in Chapter 4.

The fireball is likely to be magnetised because any turbulence in its expanding material will lead to motion of charged particles and consequent magnetic field generation. If there are any magnetic fields present, then any relativistic electrons that are present will be accelerated into spiral trajectories and emit synchrotron radiation. If the ambient magnetic fields are strong enough and the accelerated electrons are energetic enough then the synchrotron spectrum can extend to γ-ray energies and produce the observed prompt GRB emission.

5.4.4 Generating the afterglow emission

In this section we will consider the phases in the fireball evolution that are responsible for generating the GRB afterglow. Throughout the fireball's expansion it has been sweeping up material from the surrounding space. This material might be leftover remnants or debris from the celestial event that produced the GRB or it may just be material from the ambient interstellar medium. After the fireball has reached its maximum speed at $R_{\rm S}$, this continual accumulation of extra mass causes the fireball expansion to decelerate.[§] This deceleration drives a powerful relativistic shock called a **forward shock** into the fireball's surroundings, which delivers the energy that produces the afterglow emission. The afterglow photons are generated by the same mechanisms that produced the prompt γ-ray emission. The forward shock extracts the fireball's bulk kinetic energy and transfers it to electrons in the fireball, which accelerate to relativistic speeds and emit synchrotron radiation at γ-ray wavelengths.

During the prompt-emission phase, the deceleration is relatively mild and the emission from the forward shock is negligible compared to that produced by the internal shocks within the fireball. This situation changes once the mass $M_{\rm ext}$ of the swept-up external material reaches an approximate threshold such that:

$$M_{\rm ext} \gtrsim \frac{M_0}{\gamma_{\rm max}} \tag{5.21}$$

At this point the emission from the forward shock becomes comparable to that from the prompt emission, which is now rapidly fading.

The radius $R_{\rm dec}$ at which the inequality in Equation 5.21 is satisfied is called the **deceleration radius**, $R_{\rm dec}$. We can estimate $R_{\rm dec}$ if we assume a specific composition and radial density profile for the material being swept up. Let's consider a very simple case in which the ambient medium is pure hydrogen and has a constant density of n atoms per cubic metre. If we assume that the initial size of the fireball R_0 is negligible compared to $R_{\rm dec}$, then we can approximate the volume of material $V_{\rm dec}$ that the fireball has swept up once it has expanded to the deceleration radius. This is just

$$V_{\rm dec} = \frac{4\pi R_{\rm dec}^3}{3}$$

[§] Any material swept up before the fireball reaches $R_{\rm S}$ has negligible effect on its rate of expansion, so the expressions derived in Section 5.4.1 are valid.

We can write the swept-up mass that was contained within $V_{\rm dec}$ as

$$M_{\rm ext} = \rho\, V_{\rm dec} = m_{\rm p}\, n\, \frac{4\pi R_{\rm dec}^3}{3}$$

where ρ is the ambient medium density. Now we substitute this expression into Equation 5.21 and rearrange to find that

$$R_{\rm dec} = \left(\frac{3M_0}{4\pi\gamma_{\rm max} m_{\rm p} n}\right)^{1/3} \approx \left(\frac{3E_0}{4\pi\gamma_{\rm max}^2 m_{\rm p} n c^2}\right)^{1/3} \tag{5.22}$$

where we have used Equation 5.17 to replace M_0 with E_0. If $\gamma_{\rm max} \gg 1$ then its average expansion speed is approximately c and the fireball reaches the deceleration radius in a time

$$t_{\rm dec} \approx \frac{R_{\rm dec}}{c}$$

However, Equation 5.14 tells us that that an observer who sees the fireball expanding relativistically towards them with a bulk Lorentz factor $\sim\gamma_{\rm max}$ will measure a shorter interval

$$t_{\rm dec,obs} \approx \frac{R_{\rm dec}}{2\gamma_{\rm max}^2 c} \tag{5.23}$$

Exercise 5.5

Consider a GRB fireball that starts expanding with an initial energy $E_0 = 10^{44}$ J into an interstellar medium containing one atom of hydrogen per cubic centimetre. Estimate the deceleration timescale $t_{\rm dec,obs}$ that would be measured by a distant observer who sees the fireball expanding with a maximum Lorentz factor $\gamma_{\rm max} = 200$.

To observers witnessing a typical GRB from Earth, the fireball appears to reach its deceleration radius within a few seconds, which means that afterglow emission may start to arrive before the prompt emission has faded completely. This is consistent with the observed light curves shown in Figure 5.8, which show how the prompt emission observed by the *Swift* BAT instrument joins smoothly with the X-ray afterglow observed by the *Swift* XRT.

Spectral evolution and the normal decay phase

After passing $R_{\rm dec}$ the fireball continues to sweep up ambient material and the forward shock continues to grow. The fireball's rapidly increasing mass and ongoing radiative losses slow its expansion and γ decreases at an increasing rate.

As the forward shock slows down, the magnetic field strength and the typical electron energy in its vicinity are expected to decrease, which affects the spectrum of the ongoing synchrotron emission. As the synchrotron spectrum evolves, the standard theoretical expectation is that the observed γ-ray, X-ray and optical fluxes should all decay as power-law

functions of time, which is broadly consistent with the behaviour of observed GRB afterglows.

Refreshed shocks and the plateau phase

The expectation that flux should decrease following a decaying power-law trend matches the light-curve evolution that is observed during the normal decay phase of a GRB afterglow. However, it does not explain the constant or even rising X-ray fluxes that are observed during the plateau phase. To produce a flat light curve, some mechanism is needed that can maintain the forward shock velocity and replenish the energy of the electrons in its vicinity. The most popular explanation for how this happens is called the *refreshed shock model*; it is related to the internal shock model for prompt GRB emission that you learned about in Section 5.4.3.

Recall that the internal shock model assumes a fireball containing multiple different regions or shells that are expanding with slightly different Lorentz factors. In the refreshed shock model, these different shells arrive at the forward shock separately with different time delays and so deposit their combined energy over an extended period of time. Detailed modelling has shown that this mechanism can explain the plateau in a GRB afterglow light curve. The plateau phase ends when the innermost expanding shell reaches the forward shock region and the flux begins to fall as the light curve enters the normal decay phase.

Achromatic breaks and the post jet-break phase

In Section 5.2.2 you saw that some GRB afterglow light curves exhibit achromatic breaks that are characterised by abrupt changes in the rates at which their fluxes decay. These abrupt changes occur simultaneously across a wide range of frequencies. If they are observed, these breaks typically occur between the normal decay phase and the post jet-break phase. In this section we will see that achromatic breaks and the flux evolution during the post jet-break phase both provide observational evidence that GRB fireballs are not spherical and GRB emission is not isotropic.

Before we discuss the evidence that supports it, we should consider how a non-spherical geometry would modify any of the expressions that we have derived in this chapter. In particular, we will consider a *jet* geometry in which the fireball forms a biconical outflow, like the one shown in Figure 5.11.

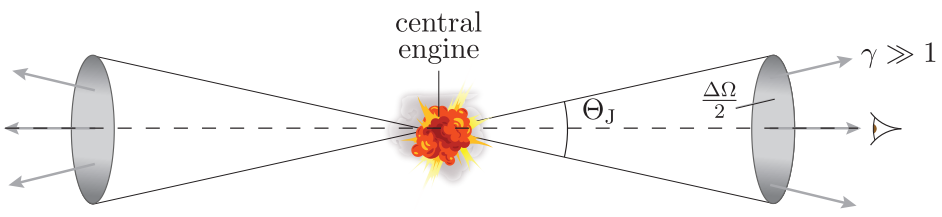


Figure 5.11 Schematic illustration of GRB with a biconical jet geometry. The mathematical symbols are discussed in the text.

If the opening angle of each conical jet is Θ_J, then the GRB energy is channelled into a solid angle

$$\Delta\Omega = 4\pi\,[1 - \cos(\Theta_J/2)] < 4\pi$$

In the following discussion, we assume that Θ_J is expressed in radians and therefore that $\Delta\Omega$ has units of steradians. Note that $\Delta\Omega$ is always less than the solid angle subtended by the surface of a sphere, which is equal to 4π. Detailed modelling that is beyond the scope of this module predicts that as long as $\gamma \gg 1$ and $\Theta_J > 2/\gamma$, the material flowing along the jets behaves exactly as it would in a relativistically expanding sphere. Therefore, the most immediate consequence of such a geometry is that the energy output we should infer based on the observed luminosity is greatly reduced.

Now, if we use the symbol E_{iso} to denote the energy we would infer from the observed GRB luminosity if we assume isotropic emission, then the true energy output E_0 can be written as:

$$E_0 = \frac{\Delta\Omega}{4\pi} E_{iso} = [1 - \cos(\Theta_J/2)]\, E_{iso}$$

If the jet opening angle is small, such that $\Theta_J \ll 1$, then we can use the Taylor series expansion $\cos x \approx 1 - x^2/2$ to write:

$$\frac{E_0}{E_{iso}} = [1 - \cos(\Theta_J/2)] \approx \left[1 - \left(1 - \Theta_J^2/8\right)\right] \approx \frac{\Theta_J^2}{8} \tag{5.24}$$

Exercise 5.6

In Section 5.2.1 you learned that the assumption of isotropic γ-ray emission implies total GRB energy outputs $E_{iso} \sim 10^{44}$ J.

(a) Using this value, calculate a new estimate for the energy output of a typical GRB assuming that its γ-ray emission is *not* isotropic, but confined within biconical jets with opening angle $\Theta_J = 6°$.

(b) By assuming isotropic γ-ray emission, so that $E_0 = E_{iso}$, a GRB's saturation and deceleration radii have been estimated to be $R_{S,iso} = 1.4 \times 10^{11}$ m and $R_{dec,iso} = 1.6 \times 10^{16}$ m, respectively. Assuming that the GRB is expanding into an ambient medium consisting of pure hydrogen that has a constant density of atoms per cubic metre, calculate new estimates for R_S and R_{dec} assuming the jet geometry specified in part (a).

To explain the observation of achromatic breaks in GRB afterglow light curves, we will consider the combined effect of relativistic beaming and a biconical jet geometry. Figure 5.12 shows how afterglow emission from points within the jet is concentrated into narrow cones orientated in the direction of the expanding fireball's bulk relativistic motion. Outside of these cones the γ-ray emission is heavily suppressed, so distant observers would not be able to detect it.

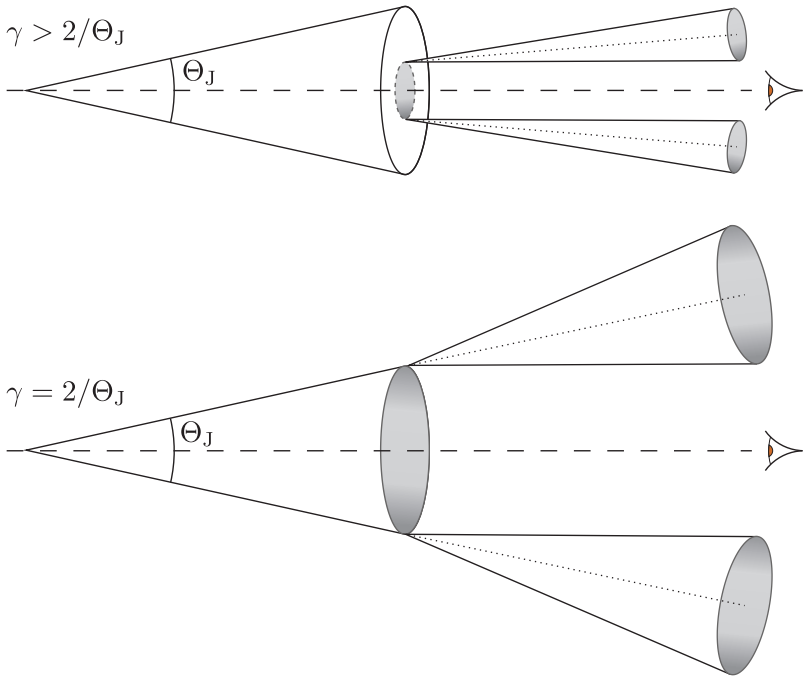


Figure 5.12 Geometrical interpretation of the jet break. In the top panel, the jet bulk Lorentz factor $\gamma > 2/\Theta_J$ and distant observers only see a small part of the jet surface area. In the bottom panel, $\gamma = 2/\Theta_J$ and the observers see the whole jet.

As the jet slows down, the opening angles of the beaming cones get wider. The achromatic break is observed at the point in time when $\gamma \sim 2/\Theta_J$. Figure 5.12 shows that once $\gamma \leq 2/\Theta_J$ then the whole jet is visible to the observer. During this phase, any observed decrease in the afterglow brightness represents changes in the physical properties of the forward shock as it slows down.

At earlier times, when $\gamma > 2/\Theta_J$, the distant observers can only detect bright emission from a small, but growing, part of the jet's overall cross-section. During this phase, the observed flux decreases more slowly because the physically driven drop in intrinsic brightness is partially compensated for by the simultaneous increase in visible luminous area.

When achromatic breaks are observed, the time that they occur can be used together with models for the evolution of γ to estimate the jet opening angle Θ_J. For the population of GRBs that have observed achromatic breaks, typical values of Θ_J range between 1 and 10 degrees, with brighter bursts like GRB 090618 tending to have narrower, more collimated jets.

5.5 Unveiling the GRB central engines

In the previous section you learned that the observed properties of GRBs can be explained in terms of synchrotron emission from collimated ultra-relativistic jets that are aligned close to the observer's line of sight and driven by a hidden central engine. In this section we will finally discuss the celestial progenitor objects that are thought to power these jets and you will see how recent astrophysical observations have helped to confirm long-standing theories about the origins of GRBs.

In Section 5.2.1 you saw that GRBs can be classified into two categories based primarily on their duration. The observed properties of long and short GRBs are sufficiently different that astronomers quickly proposed that they originate from two different types of progenitor. The theory that long and short GRBs represent physically distinct populations is reinforced by the fact that they are found in markedly different astrophysical environments.

In Section 5.2.3 you learned that long GRBs tend to occur in sites of active star formation, particularly those within irregular metal-poor galaxies. In this respect, the observed locations of long GRBs are similar to those of Type II supernovae. A Type II supernova happens when the core of a massive star undergoes catastrophic gravitational collapse to form a neutron star or black hole. Such a collapse releases very large amounts of energy. In fact, the total energy released by a typical Type II supernova is only slightly less than is required to power a typical GRB, once the effects of biconical jet geometry are accounted for. It seems reasonable to assume that long GRBs represent an extreme manifestation of the physical processes that produce Type II supernovae; we will discuss this further in Section 5.5.1.

Unlike their long GRB counterparts, short GRB afterglows have been observed throughout many different types of galaxies, which makes their spatial distribution much more like the Type Ia supernovae that you learned about in *Cosmology* Chapter 5. Each Type Ia supernova is the result of a long-lasting interaction between a compact white dwarf (WD) and another star. There is now substantial observational evidence that short GRBs result from the interaction and eventual coalescence of two compact objects. We will discuss this evidence, and the candidate progenitors for short GRBs that it implies, in Section 5.5.2.

5.5.1 Long GRBs and hypernovae

When the most massive stars exhaust their fuel for nuclear fusion, their cores can undergo catastrophic collapse. For stars with masses exceeding about $25\,M_\odot$, the end state of this collapse is thought to be a stellar-mass black hole. The gravitational collapse releases more than 10^{46} J of potential energy and although a large fraction of this energy is lost in an intense burst of neutrinos, what remains is still enough to power a typical Type II supernova. The central engines of long GRBs are believed to be more extreme analogues of core-collapse supernovae called **hypernovae**.

In Type II supernovae, around 10^{44} J of energy couples to about $1\,M_\odot$ of matter. This unbinds the whole stellar envelope and drives a quasi-spherical explosion that expands into the surrounding space at speeds $\sim 10^4\,\mathrm{km\,s^{-1}}$. In a hypernova, a similar amount of energy is imparted to at most $\sim 10^{-5}\,M_\odot$ of matter[‖] which we have established must form a highly collimated, ultra-relativistic jet.

In Chapter 4 you learned how the jets of AGNs may be powered by a spinning black hole twisting magnetic field lines that are generated by a surrounding accretion flow. A similar mechanism may drive the jets in hypernovae, but in this case the accretion disk and a surrounding torus that feeds it must both form inside the envelope of the collapsing star. To maintain a stable accretion flow in such an extreme environment the angular momentum of the inflowing material must be very high, which implies that the collapsing star must be rotating quite rapidly when its fuel runs out.

Wolf–Rayet stars are both massive and rapidly rotating, which led them to be considered as plausible hypernova progenitors. However, Wolf–Rayet stars often produce powerful equatorial winds that can carry away angular momentum and thus slow their rotation substantially before they run out of fuel. This would inhibit the maintenance of an internal accretion flow so it may be that other types of massive stars are the real sources of long GRBs. The angular momentum of these stars could be increased towards the ends of their lives via tidal interaction with a binary companion.

Even if a stable accretion flow forms within a collapsing star, it is not immediately clear whether any jet that is formed can bore a channel through the surviving stellar atmosphere while maintaining an ultra-relativistic speed that would allow distant observers to detect the jet emission as a GRB. To address this uncertainty astronomers run detailed computer simulations that model the formation of the jet and its interaction with its surroundings.

[‖] In Section 5.4.2 you saw that transparency arguments constrain the baryon load of GRBs to be at between $\sim 10^{-7}\,M_\odot$ and $\sim 5 \times 10^{-5}\,M_\odot$.

Results from one such simulation are illustrated in Figure 5.13 and show that the jet can form a turbulent *cocoon* surrounding an ultra-relativistic core with $\beta \equiv v/c \approx 1$ that escapes the stellar cocoon.

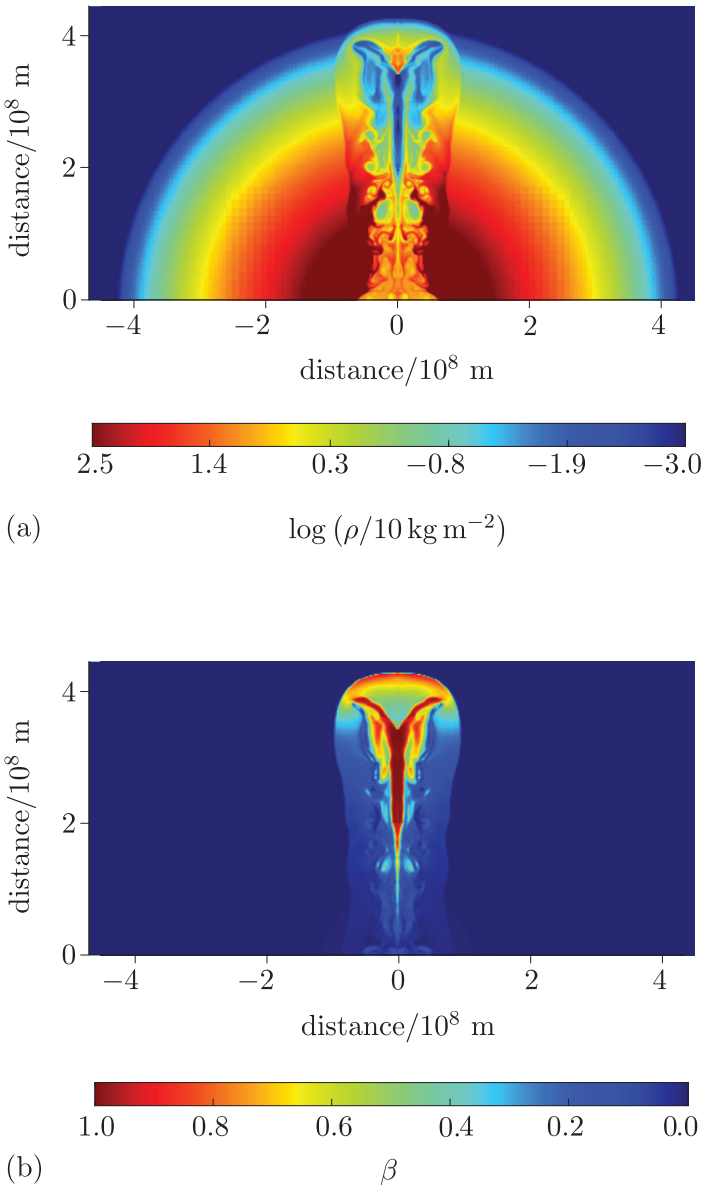


Figure 5.13 A slice through a three-dimensional simulation of a hypernova jet breaking through the atmosphere of its parent star. (a) The log-density of material in the jet and the stellar envelope. (b) The velocity of the jet material in units of the speed of light.

Even with the wealth of observational data provided by satellites like *Swift*, *Fermi* and the *Hubble Space Telescope* (*HST*), and detailed theoretical models run using powerful supercomputers, the precise origins of long GRBs remains somewhat mysterious. However, in the next section you will learn how groundbreaking discoveries made since 2016 have all-but-confirmed the previously hidden nature of short GRBs.

5.5.2 Short GRBs and compact binaries

The lives of stars with masses above $\sim 9\,M_\odot$ typically end in core-collapse supernovae that leave behind a massive and extremely dense stellar remnant. If the stellar mass exceeds $\sim 20\,M_\odot$, then this remnant is likely to be a black hole (BH); otherwise, it will be a neutron star (NS). Occasionally, two of these compact remnants form in a binary system; it is these compact binaries that are now known to be the progenitors of short GRBs.

As you learned in *Cosmology* Chapter 3, moving masses can generate gravitational waves. In a compact binary, these waves carry energy away from the system; as a result, its orbital separation decreases over time until eventually the two remnants collide and coalesce. The events before and after this coalescence are illustrated schematically in Figure 5.14.

If one or both of the remnants is a neutron star, then they can be disrupted by tidal interactions as they approach their binary companion. Some of the material that is tidally stripped from the neutron stars' surfaces forms a debris torus in the plane of the binary orbit. When the two remnants finally merge, the result is a rapidly rotating black hole.

Subsequent accretion of material from the debris torus onto the black hole is expected to launch an ultra-relativistic jet and power the short GRB emission. The compact binary merger model implies that both long and short GRBs are powered by accretion onto a newly formed, stellar-mass black hole. What differs are the overall mass of material that is accreted and the environment in which that accretion occurs.

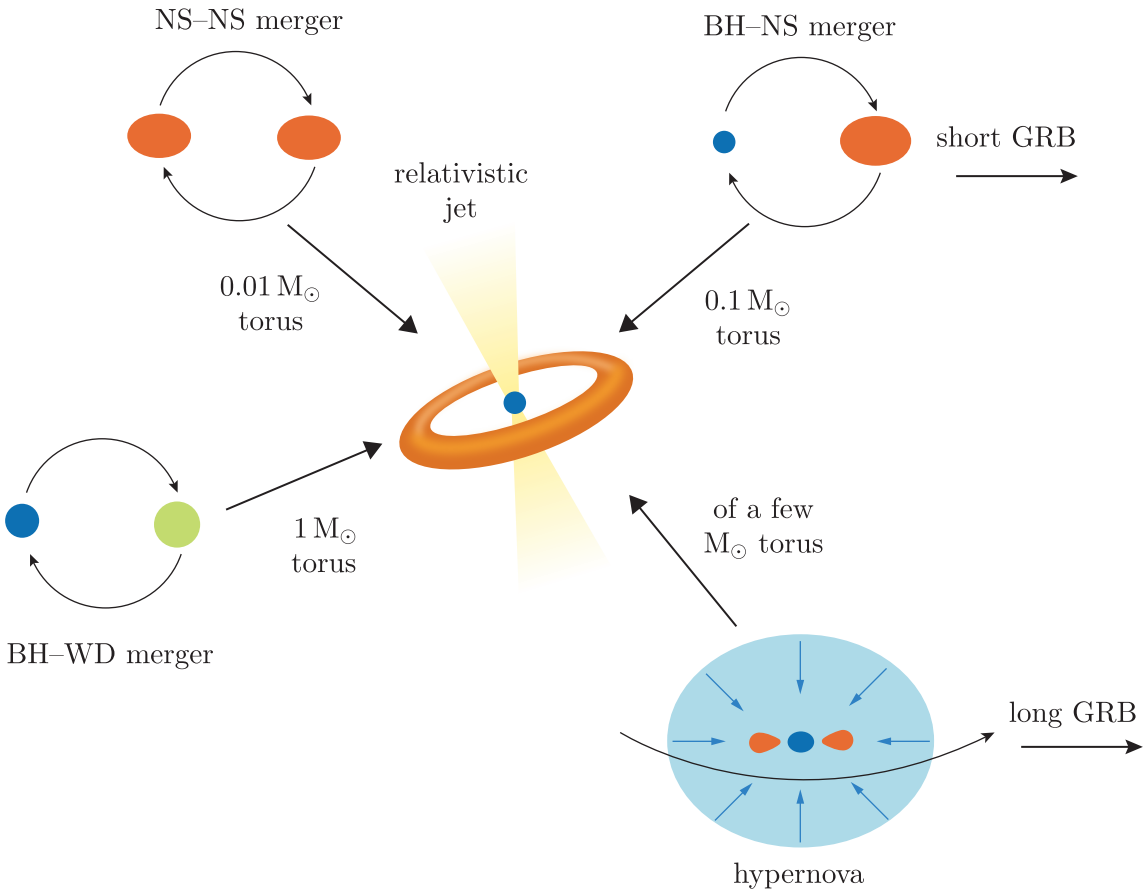


Figure 5.14 Possible GRB progenitors. Short GRBs result from the merger of two compact objects in a binary system. Long GRBs result from the collapse of a massive rotating star at the end of its life. The acronym WD stands for 'white dwarf'.

Remarkably, after several decades of uncertainty about the origins of short GRBs, there is now very strong observational evidence that the compact binary merger model is correct. On 17 August 2017, the Advanced LIGO (hereafter shortened to LIGO) and Virgo gravitational wave interferometers detected the gravitational waves that had been emitted in the final few seconds before the merger of a compact binary containing two neutron stars. The distinctive signal that LIGO detected is illustrated in the bottom panel of Figure 5.15. The intensity of the gravitational wave emission, and therefore also the rate of orbital decay, increased as the two neutron stars approach each other. At the same time, the frequency of the gravitational waves increased as the orbital period of the binary decreased. The signal stopped abruptly at the instant the two remnants merged.

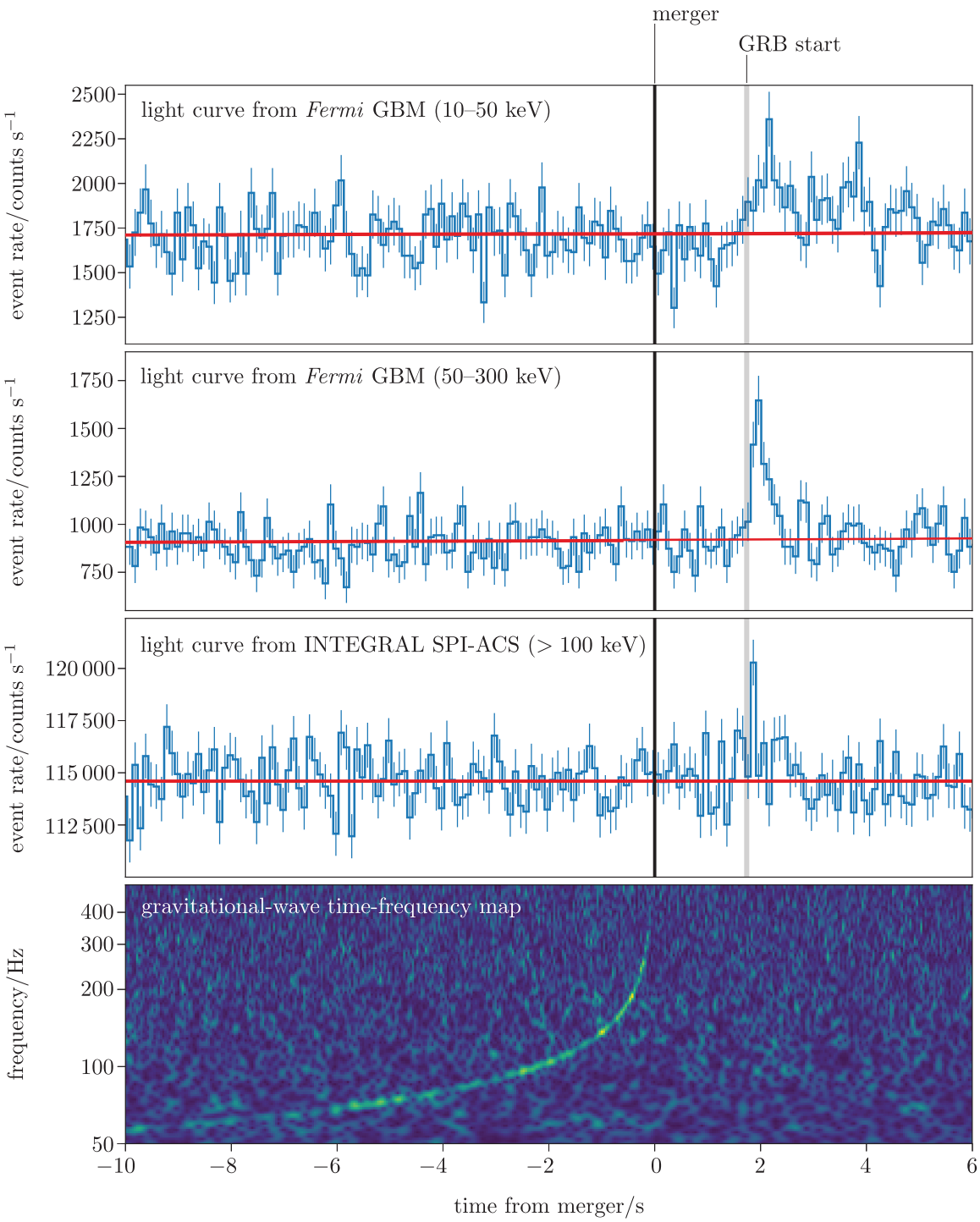


Figure 5.15 The first observational evidence that short GRBs are powered by merging compact binary systems. The top three panels show the γ-ray light curves observed by the *Fermi* GBM and the SPI-ACS instrument aboard the *INTEGRAL* X-ray telescope. The bottom panel shows the signal recorded by the LIGO gravitational wave interferometer. As the two neutron stars in the compact binary spiral closer together, the strength of the gravitational wave signal increases along with its frequency. The gravitational wave signal ends abruptly when the two neutron stars collide and coalesce. The prompt emission from the GRB begins 1.74 seconds later.

Just 1.74 seconds later, the *Fermi* GBM detected the short GRB signal illustrated in the top two panels of Figure 5.15. The burst was also detected by the *INTEGRAL* satellite, which provided an independent verification of the γ-ray signal; the corresponding light curve is shown in the third panel of Figure 5.15. The temporal coincidence between all three signals is obvious but crucially, LIGO *Fermi* and *INTEGRAL* also agreed about the direction on the sky that those signals arrived from. The gravitational wave signal was assigned the identifier GW 170817 and the near-simultaneous GRB was named GRB 170817A.

Like its predecessor, the *CGRO* BATSE, the *Fermi* GBM has coarse spatial resolution and the same is true of the LIGO interferometer. Unfortunately, the *Swift* satellite was not able to point towards the GRB location until around 15 hours after the initial γ-ray detection. The first optical detection was made 11 hours later by the Swope Telescope at Las Campanas Observatory in Chile: it located a fading optical afterglow in the outskirts of a galaxy called NGC 4993, which is about 40 Mpc from Earth. The initial Swope Telescope detection was quickly confirmed by 70 other ground- and space-based telescopes, spanning frequencies from the γ-ray to radio bands. As an example, Figure 5.16 shows *HST* observations starting about 1 day after the initial burst that monitored the fading ultraviolet, optical and near-infrared afterglow over 6 subsequent days.

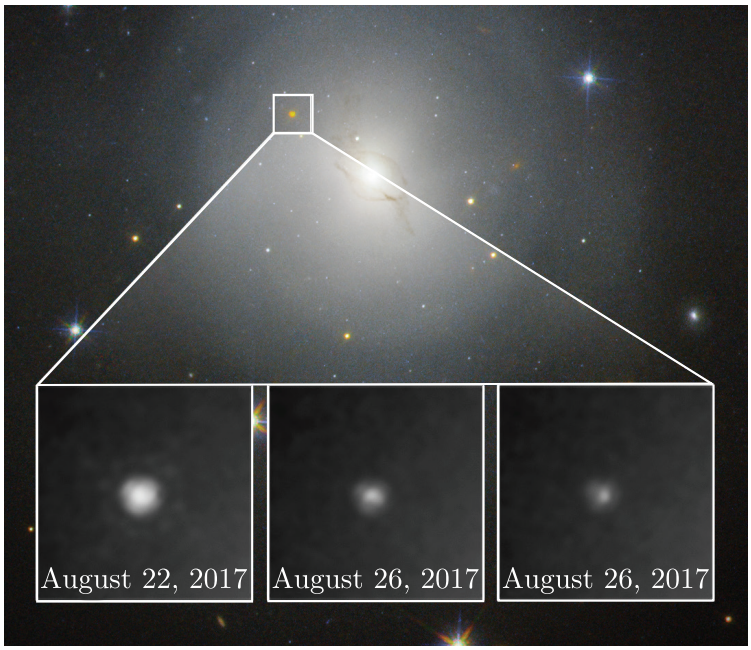


Figure 5.16 *HST* images showing the fading optical afterglow following the detection of GRB 170817A by the *Fermi* GBM.

The enormous multi-wavelength observing campaign that followed the detection of GW 170817 has provided very strong evidence that at least some short GRBs can be the result of merging neutron stars in a compact

binary system.[¶] As well as the GRB emission, the violence of the merger event itself is believed to produce a rapidly expanding and quasi-spherical explosion called a **kilonova**, which is responsible for some of the observed afterglow emission. Spectroscopic observations identified up to $10\,M_\oplus$ of heavy elements like gold and platinum among the kilonova ejecta from GRB 170817A and it is possible that kilonovae are one of the main sources of these elements in the Universe.

Now that we know compact binary mergers are responsible for at least some of the observed short GRBs, we can explain some of the observational properties that distinguish them from long GRBs. Let's start with their prompt spectra. In Section 5.2.1 you saw that short GRBs typically emit a larger proportion of high-energy photons than long GRBs. This could be explained by the fact that the jets of long GRBs form within the envelopes of their parent stars and so they can accumulate a large baryon load as they tunnel out of the stellar atmospheres. For example, massive stars tend to emit strong stellar winds throughout their lives and especially as their fuel starts to run out. This can substantially increase the density of the interstellar medium (ISM) surrounding the star and the jet may sweep up more baryons even after it escapes the stellar envelope.

In contrast, the debris surrounding the newly formed black hole following a compact binary merger is likely to be more diffuse than a stellar atmosphere or the ISM surrounding a massive star, so the baryon load of short GRB jets may be much lower. Equation 5.17 tells us that a fireball or jet with a larger baryon load will reach a lower maximum Lorentz factor than one with a lower baryon load, for the same initial energy input. If the jets of long GRBs have a larger baryon load than those of short GRBs, then their jets will tend to be less relativistic. A lower value of $\gamma_{\max}$ implies that the observed γ-ray spectrum will be less strongly Doppler blueshifted and so long GRBs will appear to emit proportionally fewer high-energy γ-rays, on average, as is observed.

The difference between the observed rates and distances to long and short GRBs can also be explained in terms of their different progenitors. Once a compact binary forms, it can take several billion years for its orbit to decay. The compact binary that produced GW 170817 was estimated to have formed about 7 billion years before its two neutron stars finally merged. This very long time lag between the formation of a compact binary system and the eventual GRB may explain the distribution of short GRB redshifts plotted in Figure 5.6, which shows that short GRBs are found predominantly in nearby galaxies. Light from more distant galaxies was emitted when the age of the Universe was smaller than the combined time taken for the first compact binary systems to form and their subsequent orbital decay, so we should not expect those galaxies to host short GRBs. In contrast, long GRBs likely happen when short-lived

[¶]Note that the possibility remains that only a subset of short GRBs result from compact binary mergers and other as-yet-unknown progenitors are responsible for the rest.

massive stars end their lives. This means that long GRBs may have been happening since the earliest stars formed, lived and then died. Accordingly, we would expect to observe long GRBs at much larger distances and Figure 5.6 confirms that this is indeed the case. Figure 5.5 shows that we observe many more long GRBs than short GRBs. This difference arises from a combination of two effects. Firstly, the theoretical rate of occurrence of long GRBs is expected to be intrinsically higher than that of short GRBs. Secondly, we observe long GRBs out to larger distances, so we are probing a much larger volume of the Universe when counting them. Therefore, even if the occurrence rates of long and short GRBs were similar, we would still expect to observe many more long GRBs.

5.6 Summary of Chapter 5

- GRBs are extremely powerful cosmic explosions with inferred γ-ray luminosities in the range 10^{42}–10^{44} W.
- The emission from GRBs is divided into two phases called the **prompt-emission phase** and the **afterglow**.
- The duration of the prompt-emission phase varies widely between bursts. The shortest GRBs have prompt emission lasting less than a second, while the longest can persist for a few tens of minutes.
- The observed duration of the prompt emission is used to divide GRBs into two populations. Those with durations longer than 2 seconds are called long GRBs and those with durations shorter than 2 seconds are called short GRBs.
- During the prompt-emission phase the γ-ray emission from GRBs can briefly outshine all other γ-ray sources in the sky and often varies on millisecond timescales.
- The fact that GRB prompt-emission light curves vary on such short timescales is evidence that the γ-ray emission comes from a very small region of space.
- The spectrum of GRB prompt emission is well modelled by a broken power-law function:

$$N_\nu \, \mathrm{d}\nu \propto \begin{cases} \left(\dfrac{\nu}{\nu_\mathrm{p}}\right)^{\alpha} \mathrm{d}\nu & \text{if } \nu \leq \nu_\mathrm{p} \\ \left(\dfrac{\nu}{\nu_\mathrm{p}}\right)^{\beta} \mathrm{d}\nu & \text{if } \nu > \nu_\mathrm{p} \end{cases} \qquad \text{(Eqn 5.2)}$$

  Observed prompt γ-ray energies range from a few thousand to a few billion electronvolts.
- The fact that we observe such high-energy γ-rays from GRBs is strong evidence that the photons are generated in a relativistic outflow with bulk Lorentz factors that could be as high as $\gamma \sim 1000$.

- If the outflows were not highly relativistic, the compact sizes and incredible luminosities of GRBs would imply that any γ-ray photons with energies larger than the electron rest-mass energy ($m_e c^2$) collide with each other to produce electron–positron pairs and we would never observe them.
- Before the relativistic nature of GRBs was realised, this apparent contradiction was referred to as the **compactness problem**.
- During the afterglow phase both the brightness of the GRB and the typical energies of the photons it emits decrease over time. During the early afterglow phase the GRB spectrum is dominated by X-ray photons. At later times the dominant emission shifts to the optical, infrared and sometimes radio bands.
- The observed flux variation during the afterglow is quite well modelled as a series of connected power-law segments. A canonical model for afterglow light curves has been defined, which comprises four power-law segments. However, only about half of all GRB afterglow light curves exhibit all four of these canonical segments.
- Astronomers have developed a physical model called the **fireball model** that predicts the observational characteristics of GRBs remarkably well. The fireball model describes the GRB as a relativistically expanding plasma. It defines a set of critical radii at which the physical properties and photon-emission mechanisms in the plasma change and lead to changes in the observed prompt emission or afterglow. The critical radii that were discussed in this chapter, from smallest to largest, are:
  - The **saturation radius** at with the expanding fireball reaches its maximum Lorenz factor, $\gamma_{\max}$. It is defined in terms of the fireball's initial mass (M_0) energy (E_0) and radius (R_0):

$$R_S = R_0\gamma_{\max} = R_0\left(1 + \frac{E_0}{M_0 c^2}\right) \approx R_0\frac{E_0}{M_0 c^2} \qquad \text{(Eqn. 5.18)}$$

  - The **dissipation radius** at which emission of the observed prompt γ-rays begins. It can be approximately defined in terms of the observed variability timescale of the prompt emission, Δt:

$$R_{\rm dis} \approx 2\gamma_{\max}^2 c\,\Delta t \qquad \text{(Eqn. 5.19)}$$

  - The **deceleration radius** ($R_{\rm dec}$) at which fireball begins to decelerate, driving a **forward shock** into the surrounding medium, which generates the observed afterglow emission. The value of $R_{\rm dec}$ depends on the properties of the surrounding medium. Assuming that medium is hydrogen gas with uniform number density $n\,\mathrm{m}^{-3}$, it can be approximately calculated using:

$$R_{\rm dec} =\approx \left(\frac{3E_0}{4\pi\gamma_{\max}^2 m_p n c^2}\right)^{1/3} \qquad \text{(Eqn. 5.22)}$$

- Photon emission during the prompt and afterglow phases is attributed to synchrotron radiation emitted by non-thermal populations of electrons with Lorentz factors (or equivalently, energies) that exhibit power-law distributions.

$$N(\gamma_e)\,\mathrm{d}\gamma_e \propto \gamma_e^{-p}\,\mathrm{d}\gamma_e \qquad \text{(Eqn. 5.20)}$$

- The observation of **achromatic** breaks in the afterglow light curves of GRBs provides strong observational evidence that GRB explosions are not isotropic, but instead form a bipolar jet structure.
- The physical processes described by the fireball model do not depend on whether an isotropic or jet-like geometry are assumed. However, the overall energy output of the GRB that we would infer from its observed photon emission is reduced in the case of a jet geometry and this shifts the critical radii that the fireball model defines.
- Both long and short GRBs are thought to signal the birth of a new stellar-mass black hole.
- Short GRBs are almost certainly associated with the merging of two neutron stars, resulting in a powerful explosion called a **kilonova**. The first strong evidence for this scenario came from the observation of a characteristic gravitational wave signal from GRB 170817A a few seconds before the first γ-rays were detected from it.
- Long GRBs are thought to be extremely energetic analogues of core-collapse supernovae. These events are called **hypernovae** and are thought to be associated with the core collapse of very massive rapidly rotating stars.

Solutions to exercises

Solution to Exercise 1.1

(a) With the provided luminosity, the ionising photon production rate is given by Equation 1.2:

$$\dot{N}_\gamma \approx 3 \times 10^{56}\,\mathrm{s}^{-1} \times \frac{10^{39}\,\mathrm{W}}{10^{13} \times 3.83 \times 10^{26}\,\mathrm{W}} \approx 7.833 \times 10^{55}\,\mathrm{s}^{-1}$$

For an escape fraction $f_{\mathrm{esc}} \sim 1$, the number of ionising photons needed to reionise a co-moving $1000\,\mathrm{Mpc}^3$ region is simply equivalent to the number of baryons in that volume (which was calculated in Example 1.1), so $N_\gamma = 7.4 \times 10^{69}$. Therefore, the number of quasars needed to reionise the co-moving $1000\,\mathrm{Mpc}^3$ volume over 600 My (or $\sim 1.89 \times 10^{16}$ s), is:

$$N_{\mathrm{Q}} = \frac{N_\gamma}{\dot{N}_\gamma t} = \frac{7.4 \times 10^{69}}{7.833 \times 10^{55}\,\mathrm{s}^{-1} \times 1.89 \times 10^{16}\,\mathrm{s}} \approx 0.0050$$

(b) The required number density is approximately 0.005 quasars per co-moving $1000\,\mathrm{Mpc}^3$, or $5 \times 10^{-6}\,\mathrm{Mpc}^{-3}$.

Solution to Exercise 1.2

(a) The number density of quasars at $z \approx 7.5$ was relatively low – a typical co-moving volume of $1\,\mathrm{Gpc}^3$ might contain only a few quasars. Quasar activity increased over time from $z \approx 7.5$ to $z \approx 2$, which was the time at which accretion onto black holes, and consequent emission, peaked. Quasar density has declined since that time, so that the number density of quasars in the present-day Universe starts to approach that at $z \approx 7.5$.

(b) We can use Equation 1.3 to relate the quasar number density at $z = 8$ to one of the measured data points (e.g. $z = 6$):

$$\frac{n_{\mathrm{Q}}(z=8)}{n_{\mathrm{Q}}(z=6)} = \frac{10^{-0.78 \times 8}}{10^{-0.78 \times 6}}$$

and so

$$n_{\mathrm{Q}}(z=8) = n_{\mathrm{Q}}(z=6) \times 10^{-1.56}$$

Reading off from the plot gives $n_{\mathrm{Q}}(z=6) \approx 20$–$30\,\mathrm{Gpc}^{-3}$, which means that $n_{\mathrm{Q}}(z=8) \approx 0.6$–$0.8\,\mathrm{Gpc}^{-3}$.

(c) The number density of luminous quasars needed for reionisation was estimated in Exercise 1.1 to be about $5 \times 10^{-6}\,\mathrm{Mpc}^{-3}$, and so to compare the numbers we need to convert our estimate from part (b) into the same units. As $1\,\mathrm{Gpc}^3 = 10^9\,\mathrm{Mpc}^3$, our estimate from part (b) converts to a value of

$$n(z=8) \approx (6\text{–}8) \times 10^{-10}\,\mathrm{Mpc}^{-3}$$

The observed population of quasars therefore appears to be much too small at the relevant redshifts to make a major contribution to reionisation.

Solution to Exercise 1.3

We can predict M_{BH} by substituting the tabulated bulge masses into Equation 1.10. So for galaxy A:

$$\frac{M_{\mathrm{BH}}}{10^9\,\mathrm{M}_\odot} = 0.49 \left(\frac{4.4 \times 10^9\,\mathrm{M}_\odot}{10^{11}\,\mathrm{M}_\odot} \right)^{1.2}$$

and so

$$M_{\mathrm{BH}} = 1.2 \times 10^7\,\mathrm{M}_\odot$$

The fractional uncertainties on each input M_{bulge} value are $\Delta M_{\mathrm{bulge}}/M_{\mathrm{bulge}}$, and so the fractional uncertainties on the black-hole mass predictions are given by:

$$\frac{\Delta M_{\mathrm{BH}}}{M_{\mathrm{BH}}} = 1.2 \left(\frac{\Delta M_{\mathrm{bulge}}}{M_{\mathrm{bulge}}} \right)$$

and for galaxy A:

$$\Delta M_{\mathrm{BH}} = 1.2 \left(\frac{1.6 \times 10^9\,\mathrm{M}_\odot}{4.4 \times 10^9\,\mathrm{M}_\odot} \right) \times 1.2 \times 10^7\,\mathrm{M}_\odot = 0.5 \times 10^7\,\mathrm{M}_\odot$$

The resulting predicted values for $M_{\mathrm{BH}} \pm \Delta M_{\mathrm{BH}}$ for all three galaxies are given in Table S1.

Table S1 Predicted black-hole masses and uncertainties.

Galaxy	predicted $M_{\rm BH}/{\rm M}_\odot$	measured $M_{\rm BH}/{\rm M}_\odot$
A	$(1.2 \pm 0.5) \times 10^7$	$(6.6 \pm 0.9) \times 10^6$
B	$(1.9 \pm 0.9) \times 10^8$	$(6.0 \pm 1.4) \times 10^6$
C	$(5.7 \pm 2.9) \times 10^5$	$(1.1 \pm 0.5) \times 10^6$

Comparing the acceptable ranges for the predicted values of $\rm M_{BH}$ with the measured ranges $M_{\rm BH} \pm \Delta M_{\rm BH}$ provided in the question leads to the conclusion that the observations for galaxies A and C are consistent with the relation set out in Equation 1.10, whereas the predicted black-hole mass for galaxy B is much larger than what is observed, and so it appears to be an outlier that deviates from the relation.

Solution to Exercise 1.4

(a) If the black hole in the centre of GN-z11 originated at $z < 20$, and we consider the starting point to be a stellar remnant (i.e. the darkest grey region in Figure 1.16), then it must have followed a growth trajectory that is steeper than the Eddington-rate accretion indicated by the blue pathway in order to reach the plotted location by $z \approx 11$. In other words, the figure shows that for a single stellar remnant to grow to the mass of GN-z11 it would require super-Eddington accretion over a period of at least $\sim$100 My. This *may* be possible, but other types of black-hole seeds, with different origins, would require less challenging accretion rates to reach GN-z11's mass by this time.

(b) The blue and green tracks on the plot show that it would be fairly straightforward for the GN-z11 black hole to grow to a mass of 10^8–$10^9\,{\rm M}_\odot$ by a redshift of 6 to 7. The accretion rates required to do this would be below the Eddington rate.

Solution to Exercise 2.1

The impact parameter b in this scenario is the radius of the Sun plus the minimum distance above the surface, i.e. $b = 1.5\,{\rm R}_\odot$.

Putting numbers into Equation 2.1 gives:

$$\begin{aligned}\widehat{\alpha} &= \frac{4 \times 6.673 \times 10^{-11}\,{\rm N\,m^2\,kg^{-2}} \times 1.99 \times 10^{30}\,{\rm kg}}{1.04 \times 10^9\,{\rm m} \times (2.998 \times 10^8\,{\rm m\,s^{-1}})^2} \\ &= 5.7 \times 10^{-6}\,{\rm rad} \\ &= 1.2 \text{ arcseconds}\end{aligned}$$

This is very small compared to the angular diameter of the Sun, and close to the limit of what could be measured from Eddington's photograph (Figure 2.3).

Solution to Exercise 2.2

Equation 2.7 can be rewritten as a quadratic equation:

$$\theta^2 - \beta\theta - \theta_{\rm E}^2 = 0$$

We can now apply the quadratic formula to solve for θ:

$$\theta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where a, b and c here are the coefficients of the three terms of the quadratic, namely $a = 1$, $b = -\beta$, and $c = -\theta_{\rm E}^2$ is the right-hand constant term (involving the mass and distances; see Equation 2.8).

Therefore

$$\theta = \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4\theta_{\rm E}^2}\right)$$

Solution to Exercise 2.3

We first need to substitute known numbers into Equation 2.8 to determine $\theta_{\rm E}$ in radians, which can then be used to calculate the physical radius of the Einstein rings.

For case (a), converting the distances to units of metres gives a distance-ratio term of

$$\frac{D_{\rm LS}}{D_{\rm L} D_{\rm S}} = 5.94 \times 10^{-25}\,{\rm m^{-1}}$$

Substituting this term, and the provided mass value, into Equation 2.8 gives

$$\begin{aligned}\theta_{\rm E} &= \sqrt{\frac{4GM}{c^2} \times 5.94 \times 10^{-25}\,{\rm m^{-1}}} \\ &= 0.0010\,{\rm rad}\end{aligned}$$

and so results in an Einstein radius of

$$\begin{aligned} r_{\rm E} &= \theta_{\rm E} D_{\rm L} \\ &= 1.27 \times 10^{21}\,{\rm m} \\ &(\approx 41\,{\rm kpc}) \end{aligned}$$

Applying the same approach to case (b) gives $\theta_{\rm E} = 4.89 \times 10^{-9}$ radians, and $r_{\rm E} = 6.0 \times 10^{11}\,{\rm m}$ ($\approx 4\,{\rm AU}$).

Solution to Exercise 2.4

(a) If $r_{\rm E} = 6.0 \times 10^{11}\,{\rm m}$ in this scenario, the diameter of the Einstein ring is twice this, i.e. $1.2 \times 10^{12}\,{\rm m}$.

A star travelling at $150\,{\rm km\,s^{-1}}$ will travel across this distance in $1.2 \times 10^{12}\,{\rm m}/150\,{\rm km\,s^{-1}} \approx 93$ days.

(b) Since $\theta_{\rm E} \propto \sqrt{M}$ (see Equation 2.8), the Einstein-crossing timescale $t_{\rm E}$ will also vary according to $\sqrt{M}$.

(c) An Earth-mass object in the system would therefore have an Einstein-crossing timescale of

$$\begin{aligned} t_{\rm E} &= \sqrt{{\rm M}_\oplus/{\rm M}_\odot} \times 93\ {\rm days} \\ &= \sqrt{3 \times 10^{-6}} \times 93\ {\rm days} \\ &\approx 3.9\ {\rm hours} \end{aligned}$$

Solution to Exercise 3.1

In the case where T is constant, the temperature gradient term is zero and Equation 3.8 can be simplified to

$$M(r) = -\frac{k_{\rm B} r^2}{G\langle m\rangle} \frac{T}{\rho(r)} \frac{{\rm d}\rho}{{\rm d}r}$$

The pressure gradient term can be obtained by differentiating the given expression:

$$\frac{{\rm d}\rho}{{\rm d}r} = -2\rho_0 \left(1 + \frac{r}{r_{\rm c}}\right)^{-3} \cdot \left(\frac{1}{r_{\rm c}}\right)$$

Substituting the expressions for $\rho(r)$ and its derivative into the expression for $M(r)$ gives

$$M(r) = \left(-\frac{k_{\rm B} T r^2}{G\langle m\rangle}\right) \left(\frac{-2\rho_0 \left(1 + \frac{r}{r_{\rm c}}\right)^{-3}}{r_{\rm c}\rho_0 \left(1 + \frac{r}{r_{\rm c}}\right)^{-2}}\right)$$

which simplifies to

$$M(r) = \frac{2 k_{\rm B} T r^2}{G\langle m\rangle r_{\rm c}(1 + r/r_{\rm c})}$$

Therefore the expression does not depend on ρ_0, as required. Substituting in the given values for (a) gives $M(r < 150\,{\rm kpc}) = 6.4 \times 10^{12}\,{\rm M}_\odot$ and for (b) $M(r < 1\,{\rm Mpc}) = 4.4 \times 10^{14}\,{\rm M}_\odot$.

Solution to Exercise 3.2

The photon energy is calculated via $E = h\nu = 1.1 \times 10^{-22}\,{\rm J}$.

The particle rest mass energies are $m_{\rm e}c^2 = 8.2 \times 10^{-14}\,{\rm J}$ and $m_{\rm p}c^2 = 1.5 \times 10^{-10}\,{\rm J}$, for the electron and proton respectively.

Therefore the ICM particles have much higher energies than the photons. Since interactions will tend to be in the direction of thermal equilibrium, it is likely that the photons will gain energy in the interaction.

Solution to Exercise 3.3

Substituting the given temperature into Equation 3.10 gives $\Delta\nu/\nu = 0.0084$. Therefore the fractional change in CMB signal caused by the SZ effect is expected to be a few orders of magnitude stronger than the cosmological anisotropies in the CMB. (Peak frequency and temperature are proportional for a black body, and so fractional deviations for the CMB are the same in both quantities).

Solution to Exercise 3.4

(a) The blue (spiral) galaxies contribute most to the total mass function at the low-mass end, whereas the red (elliptical) galaxies make up the majority of the high-mass galaxies.

(b) Within the mass range 10^{10}–$10^{11}\,{\rm M}_\odot$ the mass functions for isolated and cluster galaxies have very similar values, and so elliptical galaxies of this mass range appear equally common in isolated and cluster environments.

(c) The mass function for isolated blue galaxies has higher values than for cluster galaxies across the full mass range, and so a larger proportion of

blue (spiral) galaxies are found in isolated environments than in clusters.

(d) For spiral galaxies, the mass function has a similar shape for isolated and cluster galaxies, and so the environment doesn't seem to affect what range of masses form. For the elliptical galaxies the shape is different at low mass: it appears that low-mass ellipticals are more likely to be in clusters rather than isolated environments.

Solution to Exercise 3.5

Reading off Figure 3.14, a BCG with an apparent (X-ray measured) mass deposition rate of $\sim 200\,\mathrm{M_\odot\,y^{-1}}$ has a typical star-formation rate of $\sim 2\,\mathrm{M_\odot\,y^{-1}}$ (noting that the axes of the plot are logarithmic – it is difficult to estimate precisely from the plot, but must be around a few solar masses per year). In other words, it is forming stars at a similar rate to the Milky Way but it seems that mass is being deposited in the BCG at a rate at least 100 times that of the Milky Way. In fact, The Milky Way appears to be depleting its gas supply, so that eventually there may not be enough to form new stars, whereas the BCG appears to be building up a lot of gas that is not forming stars.

Solution to Exercise 3.6

The sound speed in this cluster is calculated via Equation 3.17:

$$c_\mathrm{s} = \sqrt{\frac{5(1.381\times10^{-23}\,\mathrm{J\,K^{-1}})\times(5\times10^{7}\,\mathrm{K})}{3\langle m\rangle}}$$

$$= 1071\,\mathrm{km\,s^{-1}}$$

Using $t = d/c_\mathrm{s}$ and substituting in the given distance of 200 kpc gives an age of $t \sim 2\times10^{8}$ years.

If the radio-galaxy expansion was 2 or 3 times faster, the age would be shorter by the same factor.

Comparing with Figure 3.13, the radio-galaxy lifetime appears comparable to cooling times in the central few kpc of nearby clusters.

Solution to Exercise 3.7

As with Example 3.4, simply integrating the expression for number density of jets of different power will just lead to the density of radio galaxies of all powers. To get the total heating rate we first need to multiply by the heating rate (jet power) for an AGN of jet power Q before performing the integral. So

$$\epsilon_\mathrm{RG}(Q)\,\mathrm{d}Q = n_0\frac{Q}{Q_*}\left(\frac{Q}{Q_*}\right)^{-\beta}\mathrm{d}Q$$

$$= n_0\left(\frac{Q}{Q_*}\right)^{1-\beta}\mathrm{d}Q$$

and the total heating rate is

$$\epsilon_\mathrm{TOT} = \frac{n_0}{Q_*^{1-\beta}}\int_{Q_1}^{Q_2} Q^{1-\beta}\,\mathrm{d}Q$$

$$= \frac{n_0}{Q_*^{1-\beta}(2-\beta)}\left[Q^{2-\beta}\right]_{Q_1}^{Q_2}$$

Substituting in the given values including $Q_1 = 10^{35}\,\mathrm{W}$ and $Q_2 = 10^{38}\,\mathrm{W}$ gives $\epsilon_\mathrm{TOT} = 7.6\times10^{31}\,\mathrm{W\,Mpc^{-3}}$.

If our assumed distribution of jet powers is accurate, then the average heating rate is similar to (within 20% of) the energy loss rate discussed in Example 3.4, and so there is enough energy from AGN jets to balance X-ray cooling in the region considered.

Solution to Exercise 4.1

Substituting in the provided angles and speeds (and remembering to convert angles to units of radians) leads to the values of β_app listed in Table S2.

Table S2 Values of β_app for different θ and β combinations.

$\boldsymbol{\beta}$	$\boldsymbol{\theta = 1°}$	$\boldsymbol{\theta = 10°}$	$\boldsymbol{\theta = 25°}$
0.5	0.017	0.17	0.39
0.9	0.16	1.4	2.1
0.99	1.7	6.9	4.1

Solution to Exercise 4.2

The provided values of V/c give Lorentz factors of $\gamma = 1.15$, 3.20 and 7.09, respectively for the three speeds.

Substituting these, and the provided value of $\theta' = \pi/2$, into Equation 4.7 gives values of

$\tan\theta \approx 1.73$, 0.329 and 0.142 for (i), (ii) and (iii), respectively. Taking the inverse tan function for each case then gives (i) $\theta = 1.05$ radians (or 60°), (ii) $\theta = 0.318$ radians (18°) and (iii) $\theta = 0.142$ radians (8.1°).

Solution to Exercise 4.3

We first need to calculate the Doppler factor, $\mathcal{D}$. We first calculate the Lorentz factor, γ, for the provided value of V/c, which gives $\gamma = 1.8983$. Then substituting for γ, V/c and the provided value of $\theta_{\rm jet}$ gives $\mathcal{D} = 3.4061$. Via Equation 4.12, the emitted luminosity density, L'_ν, will be smaller than the observed value by a factor $\mathcal{D}^{3+\alpha} = 82.4$, and so $L'_\nu = 4.5 \times 10^{22}\,{\rm W\,Hz^{-1}}$.

Solution to Exercise 4.4

(a) For $\gamma = 1000$, substituting in the charge and mass of an electron gives

$$\nu_{\rm syn} \approx \frac{(1000)^2(1.602\times10^{-19}\,{\rm C})(10^{-7}\,{\rm T})}{2\pi(9.11\times10^{-31}\,{\rm kg})} = 2.8\times10^{9}\,{\rm Hz}$$

The same calculation for a proton gives $\nu_{\rm syn} = 1.5\times10^{6}\,{\rm Hz}$.

(b) An energy of $E = 500\,{\rm MeV}$ corresponds to different Lorentz factors for the electron and proton. For the electron

$$\gamma_{\rm e} = \frac{500\times10^{6}\,{\rm eV}\times1.602\times10^{-19}\,{\rm J\,eV^{-1}}}{(9.11\times10^{-31}\,{\rm kg})(2.998\times10^{8}\,{\rm m\,s^{-1}})^2} = 978$$

For the proton, $\gamma_{\rm e} \approx 0.5$. In this case, the electron is highly relativistic, but the proton is not, because of its lower mass. The corresponding synchrotron frequency for the electron is

$$\nu_{\rm syn} \approx \frac{(978)^2(1.602\times10^{-19}\,{\rm C})(10^{-7}\,{\rm T})}{2\pi(9.11\times10^{-31}\,{\rm kg})} = 2.7\times10^{9}\,{\rm Hz}$$

The proton would produce cyclotron emission at

$$\nu_{\rm g} = \frac{(1.602\times10^{-19}\,{\rm C})(10^{-7}\,{\rm T})}{2\pi(1.67\times10^{-27}\,{\rm kg})} = 1.5\,{\rm Hz}$$

Solution to Exercise 4.5

If 40% of the available energy powers the jet, then $\eta_{\rm jet} = 0.4$. The necessary accretion rate can therefore be calculated by rearranging Equation 4.21:

$$\dot{m} = \frac{Q_{\rm jet}}{\eta_{\rm jet}c^2}$$

Substituting in the provided values therefore gives $\dot{m} = 9.735\times10^{21}\,{\rm kg\,s^{-1}}$, using the conversion from watts to SI base units: $1\,{\rm W} = 1\,{\rm kg\,m^2\,s^{-3}}$.

The accretion rate in units of solar mass per year is therefore $\dot{m} \approx 0.15\,{\rm M_\odot\,y^{-1}}$.

The AGN luminosity is now straightforward to calculate from Equation 4.22 using the provided value of $\eta_{\rm rad} = 0.1$ and our calculated $\dot{m}$; it is $L_{\rm AGN} = 8.7\times10^{37}\,{\rm W}$. In fact we could have calculated this in a simpler way just by multiplying the jet power by the ratio of $\eta_{\rm rad}/\eta_{\rm jet}$.

Solution to Exercise 4.6

We first substitute the expression for $n(E)$ into the integral term of Equation 4.23 to give

$$U_{\rm e} = \int_{E_{\rm min}}^{E_{\rm max}} E n_0 E^{-p}\,{\rm d}E$$

which simplifies to

$$U_{\rm e} = \int_{E_{\rm min}}^{E_{\rm max}} n_0 E^{1-p}\,{\rm d}E$$

We can now evaluate the integral for the two cases given. For (i), where $p = 2$, the final expression is

$$U_{\rm e} = n_0\,(\ln E_{\rm max} - \ln E_{\rm min}) = n_0\ln(E_{\rm max}/E_{\rm min})$$

In situation (ii), where p takes any value except exactly 2,

$$U_{\rm e} = \frac{n_0}{(2-p)}\left(E_{\rm max}^{2-p} - E_{\rm min}^{2-p}\right)$$

Solution to Exercise 4.7

(a) The total energy contained within the radio lobes, $E_{\rm tot}$, is the total energy *density* ($U_{\rm tot}$) multiplied by the source volume. If the plasma is at equipartition then the magnetic field energy density makes up half of $U_{\rm tot}$, so we can work out $U_{\rm B} = B^2/(2\mu_0)$ and multiply by 2. Hence

$U_{\rm tot} = 1.27 \times 10^{-11}\,{\rm J\,m^{-3}}$ (where we retain some extra precision for later calculations). The volume V is given by considering a cylinder of length $2 \times 250\,{\rm kpc}$ (because there are two lobes) and radius 30 kpc, so that $V = \pi r^2 l = 4.15 \times 10^{64}\,{\rm m^3}$. Therefore $E_{\rm tot} = U_{\rm tot} V = 5.29 \times 10^{53}\,{\rm J}$.

(b) The enthalpy is obtained by adding a factor of $P_{\rm ext}V$ to $E_{\rm tot}$, so H is equal to $5.29 \times 10^{53}\,{\rm J} + 3.55 \times 10^{51}\,{\rm J} \approx 5.3 \times 10^{53}\,{\rm J}$.

(c) To produce the observed radio source, the jet power must have been sufficient to supply the energy needed for the current internal energy and the work done over the time period the jet has been active. In other words, $Q_{\rm jet}$ must be at least H/t, where t is the source age. Therefore $Q_{\rm jet}$ must be at least:

$$\frac{5.3 \times 10^{53}\,{\rm J}}{(10^8\,{\rm y} \times 3.156 \times 10^7\,{\rm s\,y^{-1}})} = 1.7 \times 10^{38}\,{\rm W}$$

(You will obtain slightly different values if you do no intermediate rounding.)

Solution to Exercise 5.1

From Figure 5.4, there is a peak (or at least a pronounced shoulder) on the short-GRB side of the vertical dashed line representing GRBs with typical durations $\sim$0.5 s, so we will use this as T_{90}.

From Figure 5.5, a typical short GRB fluence is $S_{15\text{–}150\,{\rm keV}} \sim 10^{-10}\,{\rm J\,m^{-2}}$.

These values are only approximate and you may have chosen slightly different ones.

Using our chosen values for $S_{15\text{–}150\,{\rm keV}}$ and T_{90}, together with the luminosity distance we were given and our assumption that the GRB emission is isotropic, we can estimate the luminosity using:

$$\begin{aligned} L_{15\text{–}150\,{\rm keV}} &= \frac{4\pi d_{\rm L}^2 S_{15\text{–}150\,{\rm keV}}}{T_{90}} \\ &= \frac{4\pi \times (6200\,{\rm Mpc})^2 \times 10^{-10}\,{\rm J\,m^{-2}}}{0.5\,{\rm s}} \\ &= \frac{4\pi \times (1.91 \times 10^{26})^2\,{\rm m^2} \times 10^{-10}\,{\rm J\,m^{-2}}}{0.5\,{\rm s}} \\ &\approx 9 \times 10^{43}\,{\rm W} \end{aligned}$$

Solution to Exercise 5.2

(a) To answer this part we simply use Equation 5.15, substituting the values we were given in the question for γ and $\Delta t_{\rm min}$.

$$\begin{aligned} R &\sim \gamma^2 c\,\Delta t_{\rm min} \\ &\sim 100^2 \times 3 \times 10^5\,{\rm km\,s^{-1}} \times 10^{-3}\,{\rm s} \\ &\sim 3 \times 10^6\,{\rm km} \end{aligned}$$

(b) From the table of constants, $\mathrm{R}_\odot = 6.96 \times 10^8\,{\rm m}$, so the prompt-emission region is only just over four times the size of the Sun, even allowing for relativistic effects.

Solution to Exercise 5.3

(a) To solve this part we use the formula for the saturation radius (Equation 5.18), using the values for R_0 and $\gamma_{\rm max}$ given in the question:

$$\begin{aligned} R_{\rm S} &= R_0 \gamma_{\rm max} \\ &= 200\,\mathrm{R}_\odot \\ &= 200 \times 6.96 \times 10^8\,{\rm m} \\ &= 1.4 \times 10^8\,{\rm km} \end{aligned}$$

(b) To solve this part we first rearrange Equation 5.17 to isolate M_0

$$M_0 \approx \frac{E_0}{\gamma_{\rm max} c^2}$$

Using the numerical values given in the question:

$$\begin{aligned} M_0 &\approx \frac{10^{44}\,{\rm J}}{200 \times (2.998 \times 10^8\,{\rm m\,s^{-1}})^2} \\ &\approx 5.56 \times 10^{24}\,{\rm kg} \\ &\approx \frac{5.56 \times 10^{24}\,{\rm kg}}{1.99 \times 10^{30}\,{\rm kg\,M_\odot^{-1}}} \\ &\approx 2.8 \times 10^{-6}\,\mathrm{M}_\odot \end{aligned}$$

Solution to Exercise 5.4

To solve this problem we will use Equation 5.17 with $E_0 \sim 10^{44}\,{\rm J}$.

$$\gamma_{\rm max} = 1 + \frac{E_0}{M_0 c^2}$$

The minimum value of M_0 that is consistent with observing γ-rays of energy $\gg$0.5 MeV that have a

non-thermal spectrum is $M_{0,\min} \sim 10^{-7}\,\mathrm{M}_\odot$. Using this value we evaluate the denominator in the second term:

$$\begin{aligned} M_0c^2 &= 10^{-7} \times 1.99 \times 10^{30}\,\mathrm{kg} \times (3 \times 10^8\,\mathrm{m\,s^{-1}})^2 \\ &= 1.79 \times 10^{40}\,\mathrm{J} \end{aligned}$$

Using this result, we evaluate Equation 5.17:

$$\begin{aligned} \gamma_{\max} &= 1 + \frac{E_0}{M_0c^2} = 1 + \frac{10^{44}\,\mathrm{J}}{1.79 \times 10^{40}\,\mathrm{J}} \\ &\approx 5600 \end{aligned}$$

The maximum value of M_0 that is consistent with observing γ-rays with energy $\gg 0.5\,\mathrm{MeV}$ is $M_{0,\min} \sim 5 \times 10^{-5}\,\mathrm{M}_\odot$. Again, we evaluate the denominator in the second term. This time we do so by noting that the value of M_0c^2 must be 500 times the value we calculated in the first case:

$$\begin{aligned} M_0c^2 &= 500 \times 1.79 \times 10^{40}\,\mathrm{J} \\ &= 8.96 \times 10^{42}\,\mathrm{J} \end{aligned}$$

Hence the first expression now gives

$$\begin{aligned} \gamma_{\max} &= 1 + \frac{E_0}{M_0c^2} = 1 + \frac{10^{44}\,\mathrm{J}}{8.95 \times 10^{42}\,\mathrm{J}} \\ &\approx 12 \end{aligned}$$

Solution to Exercise 5.5

To solve this problem we first combine Equations 5.22 and 5.23 to write

$$\begin{aligned} t_{\mathrm{dec,obs}} &\approx \frac{R_{\mathrm{dec}}}{2\gamma_{\max}^2 c} \\ &\approx \frac{1}{2\gamma_{\max}^2 c}\left(\frac{3E_0}{4\pi\gamma_{\max}^2 m_{\mathrm{p}} n c^2}\right)^{1/3} \\ &\approx \left(\frac{3E_0}{32\pi\gamma_{\max}^8 m_{\mathrm{p}} n c^5}\right)^{1/3} \end{aligned}$$

Now, using the values given in the question and noting that $1\,\mathrm{cm^{-3}} = 10^6\,\mathrm{m^{-3}}$ we evaluate this expression

$$\begin{aligned} t_{\mathrm{dec,obs}} &\approx \left[\frac{3 \times 10^{44}\,\mathrm{J}}{32\pi \times 200^8 \times 1.6726 \times 10^{-27}\,\mathrm{kg}} \right. \\ &\quad \left. \times \frac{1}{10^6\,\mathrm{m^{-3}} \times (2.998 \times 10^8\,\mathrm{ms^{-2}})^5}\right]^{1/3} \\ &\approx 6.6\,\mathrm{s} \end{aligned}$$

Solution to Exercise 5.6

(a) To solve this part we can use Equation 5.24. First, we need to convert the value of $\Theta_{\mathrm{J}} = 6°$ from units of degrees to radians. Combining these two steps, we find

$$\begin{aligned} E_0 &= \frac{(6\text{ degrees})^2}{8} \times 10^{44}\,\mathrm{J} = \frac{(6\pi)^2 \times 10^{44}}{8 \times 180^2}\,\mathrm{J} \\ &= 1.4 \times 10^{41}\,\mathrm{J} \end{aligned}$$

(b) To solve this part we need to calculate the factor by which the value of E_0 calculated in part (a) is smaller than E_{iso}. If we use the symbol f_{jet} to denote this factor, then we can write:

$$f_{\mathrm{jet}} = \frac{E_0}{E_{\mathrm{iso}}} = \frac{\Theta_{\mathrm{J}}^2}{8} = \frac{(6\pi)^2}{8 \times 180^2} = 0.0014$$

Now with reference to Equations 5.18 and 5.22 we can write down how R_{S} and R_{dec} depend on E_0 i.e.

$$\begin{aligned} R_{\mathrm{S}} &\propto E_0 \\ R_{\mathrm{dec}} &\propto E_0^{1/3} \end{aligned}$$

Using these proportionality relationships and introducing the symbols $R_{\mathrm{S,\,jet}}$ and $R_{\mathrm{dec,\,jet}}$ to denote estimates of the saturation and deceleration radii under the assumption of a biconical jet geometry, we can estimate the saturation radius

$$\begin{aligned} R_{\mathrm{S,\,jet}} &= f_{\mathrm{jet}} R_{\mathrm{S,iso}} = \frac{(6\pi)^2}{8 \times 180^2} \times 1.4 \times 10^{11}\,\mathrm{m} \\ &= 1.9 \times 10^8\,\mathrm{m} \end{aligned}$$

and the deceleration radius as

$$\begin{aligned} R_{\mathrm{dec,\,jet}} &= f_{\mathrm{jet}}^{1/3} R_{\mathrm{dec,iso}} \\ &= \left[\frac{(6\pi)^2}{8 \times 180^2}\right]^{1/3} \times 1.6 \times 10^{16}\,\mathrm{m} \\ &= 0.111 \times 1.6 \times 10^{16}\,\mathrm{m} \\ &= 1.78 \times 10^{15}\,\mathrm{m} \end{aligned}$$

References and acknowledgements

References

Cavagnolo, K. W. *et al.* (2009) *ACCEPT: archive of* Chandra *cluster entropy profile tables*, Abell 1795, ObsID 493. Available at: https://web.pa.msu.edu/astro/MC2/accept/clusters/493.html (Accessed: 26 January 2024).

Dressler, A. (1980) 'Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies', *Astrophysical Journal*, 236, pp. 351–365. Available at https://doi.org/ 10.1086/157753.

Dyson, F. W. *et al.* (1920) 'IX. A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of May 29, 1919', *Philosophical Transactions of the Royal Society A*, 220(571–581), pp. 291–333. Available at https://doi.org/10.1098/rsta.1920.0009.

Hardcastle, M. J. *et al.* (2003) 'Radio and X-ray observations of the jet in Centaurus A', *The Astrophysical Journal*, 593(1), pp. 169–183. Available at https//doi.org/10.1086/376519.

Ryden, B. (2017) *Introduction to cosmology*. 2nd edn. New York: Cambridge University Press.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: NASA and the Space Telescope Science Institute (STScI).

Chapter images: Figure 1.1: Robertson, B. E. (2022) 'Galaxy formation and reionization: key unknowns and expected breakthroughs by the *James Webb Space Telescope*', *Annual Review of Astronomy and Astrophysics*, 60, pp. 121–158, Annual Reviews, https://www.annualreviews.org/doi/10.1146/annurev-astro-120221-044656, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence, https://creativecommons.org/licenses/by/4.0/; Figures 1.2 and 1.4: Wise, J. H. (2019) 'An introductory review on cosmic reionization', Georgia Institute of Technology, Atlanta; Figure 1.5: Carilli, C. L. *et al.* (2006) 'Observational constraints on cosmic reionization', *Annual Review of Astronomy and Astrophysics*, 44(1), pp. 415–462, Annual Reviews, https://www.annualreviews.org/journal/astro, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence, https://creativecommons.org/licenses/by/4.0/; Figure 1.6: NASA, ESA, CSA, M. Zamani (ESA/Webb), Leah Hustak (STScI), Brant Robertson (UC Santa Cruz), S. Tacchella (Cambridge), E. Curtis-Lake (UOH), S. Carniani (Scuola Normale Superiore), JADES Collaboration; Figure 1.8: Orlitova, I. (2020) 'Starburst galaxies', written for an ERASMUS textbook, NASA ADS, https://arxiv.org/abs/2012.12378.pdf, licensed

Society, Royal Astronomical Society; Figure 2.15: Stephane Colombi, International Astronomical Union and Canada–France–Hawaii Telescope; Figure 2.17: NASA/CANUCS; Figure 2.18: NASA, ESA, CSA and T. Treu (UCLA), https://esawebb.org/images/weic2220a/, licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, https://creativecommons.org/licenses/by/4.0/; Figure 3.1a: NASA, ESA, the Hubble Heritage team (STScI/AURA), J. Blakeslee (NRC Herzberg Astrophysics Program, Dominion Astrophysical Observatory) and H. Ford (JHU); Figure 3.1b: NASA/CXC/BU/E. Blanton; optical: ESO/VLT; Figure 3.2: NASA/CXC/University of Alabama/A. Morandi *et al.*; optical: SDSS, NASA/STScI; Figure 3.3: Sanders, J. S. *et al.* (2018) 'Enrichment in the Centaurus cluster of galaxies', *Monthly Notices of the Royal Astronomical Society*, 371(3), pp. 1483–1496, Oxford University Press, Royal Astronomical Society, Wiley-Blackwell; Figure 3.4: adapted from Phillips, A. C. (1999) *The Physics of Stars*, 2nd edn, John Wiley and Sons Limited; Figure 3.5: Pointecouteau, E. *et al.* (2005) 'The structural and scaling properties of nearby galaxy clusters', *Astronomy & Astrophysics*, 435(1), EDP Sciences; Figure 3.6: Mroczkowski, T. *et al.* (2019) 'Astrophysics with the spatially and spectrally resolved Sunyaev–Zeldovich effects', *Space Science Reviews*, 215, article number 17, Springer Link; Figure 3.7a: Khatri, R. and Gaspari, M. (2016) 'Thermal SZ fluctuations in the ICM: probing turbulence and thermodynamics in Coma cluster with *Planck*', *Monthly Notices of the Royal Astronomical Society*, 463(1), Oxford University Press, Royal Astronomical Society, Wiley-Blackwell; Figure 3.7b: Plagge, T. J. *et al.* (2013) 'CARMA measurements of the Sunyaev–Zeldovich effect in RX J1347.5–1145', *Astrophysical Journal*, 770(2), IOP Publishing; Figure 3.8: Dressler, A. *et al.* (1980) 'Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies', *The Astrophysical Journal*, 236, pp. 351–365, The American Astronomical Society; Figure 3.9: Peng, Y. *et al.* (2010) 'Mass and environment as drivers of galaxy evolution in SDSS and zCOSMOS and the origin of the Schechter function', *The Astrophysical Journal*, 721(1), pp. 193–221, The American Astronomical Society; Figure 3.10: NASA, ESA, CSA, STScI; Figure 3.11: Roberts, I. D. *et al.* (2021) 'LoTSS jellyfish galaxies I. Radio tails in low redshift clusters', *Astronomy & Astrophysics*, 650, EDP Sciences, an open access article licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, https://creativecommons.org/licenses/by/4.0/, Figure 3.12: Kenneth W. Cavagnolo, images generated using SAOImage DS9, developed by the Smithsonian Astrophysical Observatory; Figure 3.13: Panagoulia, E. K. *et al.* (2013) 'A volume-limited sample of X-ray galaxy groups and clusters - I. Radial entropy and cooling time profiles', *Monthly Notices of the Royal Astronomical Society*, 438(3), pp. 2341–2354, Oxford University Press, Royal Astronomical Society, Wiley-Blackwell; Figure 3.14: McDonald, M. *et al.* (2018) 'Revisiting the cooling flow problem in galaxies, groups, and clusters of galaxies', *The Astrophysical Journal*, 858(1), p. 45, © 2018, The American Astronomical Society; Figure 3.15:

X-ray: NASA/CXC/University of Waterloo/B. McNamara; optical: NASA/ESA/STScI/University of Waterloo/B. McNamara; radio: NRAO/Ohio University/L. Birzan *et al.*; Figure 4.1: Goddi, C;, *et al.* (2019) 'First M87 Event Horizon Telescope results and the role of ALMA', *The Messenger*, 177, pp. 25–35, European Southern Observatory; Figure 4.2: Falcke, H. (2022) 'The road toward imaging a black hole: a personal perspective', *Natural Sciences*, 2(4), Wiley, https://onlinelibrary.wiley.com/doi/10.1002/ntls.20220031, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence, https://creativecommons.org/licenses/by/4.0/; Figure 4.4: image by Glenn Piner; Figure 4.6: ESA/Gaia/DPAC, https://www.gaia.ac.uk/science/edr3-acceleration-solar-system, licensed under a Creative Commons Attribution 3.0 Intergovernmental Organization (CC BY-SA 3.0 IGO) licence, https://creativecommons.org/licenses/by/3.0/igo/; Figure 4.10a: Dr Adrian Jannetta/Michael Richmond, http://spiff.rit.edu/classes/ast613/lectures/radio_i/radio_i.html, licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic (CC BY-NC-SA 2.0) licence, https://creativecommons.org/licenses/by-nc-sa/2.0/; Figure 4.10b: R. Craig Walker/NRAO, https://www.aoc.nrao.edu/ cwalker/M87/, licensed under a Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, http://creativecommons.org/licenses/by/3.0/; Figure 4.14: Takhistov, V. (2019) 'Positrons from primordial black hole microquasars and gamma-ray bursts', *Physics Letters B*, 789, pp. 538-544, Elsevier, https://www.sciencedirect.com/science/article/pii/S0370269318309742, an open access article licensed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license, https://creativecommons.org/licenses/by/4.0/; Figure 5.7: NASA/JPL-Caltech and the Hubble Heritage team (STScI/AURA); Figure 5.13: Lazzati, D. *et al.* (2012) 'Unifying the zoo of jet-driven stellar explosions', *The Astrophysical Journal*, 750(1), IOP Publishing; Figure 5.14: adapted from Zhang, W., Woosley, S. E. and MacFadyen, A. I. (2003) 'Relativistic jets in collapsars', *The Astrophysical Journal*, IOP Publishing; Figure 5.15: Abbott, B. P. *et al.* (2017) 'Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A', *The Astrophysical Journal Letters*, 848(2), The American Astronomical Society, https://iopscience.iop.org/article/10.3847/2041-8213/aa920c, licensed under a Creative Commons Attribution 3.0 Unported (CC BY 3.0) licence, https://creativecommons.org/licenses/by/3.0/; Figure 5.16: NASA and ESA: A. J. Levan (University of Warwick), N. R. Tanvir (University of Leicester), and A. Fruchter and O. Fox (STScI), released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, https://creativecommons.org/licenses/by/4.0/. 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 Stephen Serjeant.

The authors would like to thank Bonny Barkus, Kate Gibson, Mark Jones and Sheona Urquhart for useful feedback and discussions.

External assessor

Stephen Wilkins, University of Sussex.

Curriculum team

Jessica Bartlett and Shelah Surgey.

Production team

Senior project manager

Jeni Aldridge.

Editors

Jonathan Martyn, Peter Twomey, Yon-Hee Kim, Lil Davies and Jonathan Darch.

Graphics

Sha'ni Hirschy, Anna Jordan and Helen Panayi.

OU Library

James Salter.

Index

Note: **bold** page numbers indicate where terms are defined.